A Proof of Initialization

In this section, we will prove Lemma 4.1 and a corresponding lemma for asymmetric case as follows (which will be used to prove Theorem 3.3):

Lemma A.1. *Assume $\mathbf{M} \in \mathbb{R}^{d_1 \times d_2}$ is a rank k matrix with μ-incoherence, and Ω is a subset unformly i.i.d sampled from all coordinate. Let $\mathbf{U}_0\mathbf{V}_0^\top$ be the top-k SVD of $\frac{d_1 d_2}{m}\mathcal{P}_\Omega(\mathbf{M})$, where $|\Omega| = m$. Let $d = \max\{d_1, d_2\}$. Then there exists universal constant c_0, for any $m \geq c_0 \mu d k^2 \kappa^2(\mathbf{M}) \log d$, with probability at least $1 - \frac{1}{d^{10}}$, we have:*

$$\|\mathbf{M} - \mathbf{U}_0\mathbf{V}_0^\top\|_F \leq \frac{1}{20}\sigma_{\min}(\mathbf{M}),$$
$$\max_i \left\|\mathbf{e}_i^\top \mathbf{U}_0\mathbf{V}_0^\top\right\|^2 \leq \frac{10\mu k}{d_1}\|\mathbf{M}\|, \quad \max_j \left\|\mathbf{e}_j^\top \mathbf{V}_0\mathbf{U}_0^\top\right\|^2 \leq \frac{10\mu k}{d_2}\|\mathbf{M}\| \tag{6}$$

We will focus mostly on Lemma A.1, and prove Lemma 4.1 as a special case. Most of the argument of this section follows from [14]. We include here for completeness. The remaining of this section can be viewed as proving both the Frobenius norm claim and incoherence claim of Lemma A.1 seperately.

In this section, We always denote $d = \max\{d_1, d_2\}$. For simplicity, WLOG, we also assume $\|\mathbf{M}\| = 1$ in all proof. Also, when it's clear from the context, we use κ to specifically to represent $\kappa(\mathbf{M})$. Then $\sigma_{\min}(\mathbf{M}) = \frac{1}{\kappa}$. Also in the proof, we always denote $\text{SVD}(\mathbf{M}) = \mathbf{X}\mathbf{S}\mathbf{Y}^\top$, and $\text{SVD}(\mathbf{U}\mathbf{V}^\top) = \mathbf{W}_\mathbf{U}\mathbf{D}\mathbf{W}_\mathbf{V}^\top$, where $\mathbf{S}$ and $\mathbf{D}$ are $k \times k$ diagonal matrix.

A.1 Frobenius Norm of Initialization

Theorem A.2 (Matrix Bernstein [25])**.** *A finite sequence $\{\mathbf{X}_t\}$ of independent, random matrices with dimension $d_! \times d_2$. Assume that each matrix satisfies:*

$$\mathbb{E}\mathbf{X}_t = 0, \quad \textit{and} \quad \|\mathbf{X}_t\| \leq R \ \textit{almost surely}$$

Define

$$\sigma^2 = \max\{\left\|\sum_t \mathbb{E}(\mathbf{X}_t\mathbf{X}_t^\top)\right\|, \left\|\sum_t \mathbb{E}(\mathbf{X}_t^\top\mathbf{X}_t)\right\|\}$$

Then, for all $s \geq 0$,

$$\Pr(\left\|\sum_t \mathbf{X}_t\right\| \geq s) \leq (d_1 + d_2) \cdot \exp(\frac{-s^2/2}{\sigma^2 + Rs/3})$$

Lemma A.3. *Let $|\Omega| = m$, then there exists universal constant C, c_0, for any $m \geq c_0 \mu d k \log d$, with probability at least $1 - \frac{1}{d^{10}}$, we have:*

$$\left\|\mathbf{M} - \frac{d_1 d_2}{m}\mathcal{P}_\Omega(\mathbf{M})\right\| \leq C\sqrt{\frac{\mu d k \log d}{m}}$$

Proof. We know

$$\left\|\mathbf{M} - \frac{d_1 d_2}{m}\mathcal{P}_\Omega(\mathbf{M})\right\| = \frac{d_1 d_2}{m}\left\|\mathcal{P}_\Omega(\mathbf{M}) - \frac{m}{d_1 d_2}\mathbf{M}\right\|$$

and note:

$$\mathcal{P}_\Omega(\mathbf{M}) - \frac{m}{d_1 d_2}\mathbf{M} = \sum_{ij}\mathbf{M}_{ij}(Z_{ij} - \frac{m}{d_1 d_2})\mathbf{e}_i\mathbf{e}_j^\top$$

where Z_{ij} are independence Bernoulli$(m/d_1 d_2)$ random variables. Let matrix

$$\psi_{ij} = \mathbf{M}_{ij}(Z_{ij} - \frac{m}{d_1 d_2})\mathbf{e}_i\mathbf{e}_j^\top$$

By construction, we have:

$$\left\|\sum_{ij}\psi_{ij}\right\| = \left\|\mathcal{P}_\Omega(\mathbf{M}) - \frac{m}{d_1 d_2}\mathbf{M}\right\|$$

Clearly $\mathbb{E}\psi_{ij} = 0$. Let $\mathbf{X}\mathbf{S}\mathbf{Y}^\top = \text{SVD}(\mathbf{M})$, then by μ-incoherence of $\mathbf{M}$, with probability 1:

$$\|\mathbf{M}\|_\infty \le \max_{ij} |\mathbf{e}_i^\top \mathbf{X}\mathbf{S}\mathbf{Y}^\top \mathbf{e}_j| \le \|\mathbf{M}\| \frac{\mu k}{\sqrt{d_1 d_2}}$$

Also:

$$\begin{aligned}
\left\|\sum_{ij} \mathbb{E}(\psi_{ij}\psi_{ij}^\top)\right\| &= \left\|\sum_{ij} \mathbb{E}\mathbf{M}_{ij}^2 (Z_{ij} - \frac{m}{d_1 d_2})^2 \mathbf{e}_i \mathbf{e}_i^\top\right\| \le \frac{m}{d_1 d_2}(1 - \frac{m}{d_1 d_2}) \left\|\sum_{ij} \mathbf{M}_{ij}^2 \mathbf{e}_i \mathbf{e}_i^\top\right\| \\
&= \frac{m}{d_1 d_2}(1 - \frac{m}{d_1 d_2}) \max_i \sum_j \mathbf{M}_{ij}^2 \le \frac{2m}{d_1 d_2} \frac{\mu k}{d_1} \|\mathbf{M}\|^2 = \frac{2m\mu k}{d_1^2 d_2} \|\mathbf{M}\|^2 \\
\left\|\sum_{ij} \mathbb{E}(\psi_{ij}^\top\psi_{ij})\right\| &= \left\|\sum_{ij} \mathbb{E}\mathbf{M}_{ij}^2 (Z_{ij} - \frac{m}{d_1 d_2})^2 \mathbf{e}_j \mathbf{e}_j^\top\right\| \le \frac{m}{d_1 d_2}(1 - \frac{m}{d_1 d_2}) \left\|\sum_{ij} \mathbf{M}_{ij}^2 \mathbf{e}_j \mathbf{e}_j^\top\right\| \\
&= \frac{m}{d_1 d_2}(1 - \frac{m}{d_1 d_2}) \max_j \sum_i \mathbf{M}_{ij}^2 \le \frac{2m}{d_1 d_2} \frac{\mu k}{d_2} \|\mathbf{M}\|^2 = \frac{2m\mu k}{d_1 d_2^2} \|\mathbf{M}\|^2
\end{aligned}$$

Then, by matrix Bernstein (Theorem A.2), we have:

$$\Pr\left(\left\|\sum_{ij} \psi_{ij}\right\| \ge s\right) \le 2(d_1 + d_2) \cdot \exp\left(\frac{-s^2/2}{\frac{2m\mu dk}{d_1^2 d_2^2} \|\mathbf{M}\|^2 + \|\mathbf{M}\| \frac{\mu k}{3\sqrt{d_1 d_2}} s}\right)$$

That is, with probability at least $1 - \frac{1}{d^{10}}$, for some universal constant C, we have:

$$\left\|\mathcal{P}_\Omega(\mathbf{M}) - \frac{m}{d_1 d_2}\mathbf{M}\right\| \le C \|\mathbf{M}\| \cdot \max\left\{\sqrt{\frac{m\mu dk \log d}{d_1^2 d_2^2}}, \frac{\mu k \log d}{\sqrt{d_1 d_2}}\right\}$$

For $m \ge \mu dk \log d$, we finishes the proof. □

Theorem A.4. *Let $\mathbf{U}_0\mathbf{V}_0^\top$ be the top-k SVD of $\frac{d_1 d_2}{m}\mathcal{P}_\Omega(\mathbf{M})$, where $|\Omega| = m$ then there exists universal constant c_0, for any $m \ge c_0 \mu dk^2 \kappa^2 \log d$, with probability at least $1 - \frac{1}{d^{10}}$, we have:*

$$\left\|\mathbf{M} - \mathbf{U}_0\mathbf{V}_0^\top\right\|_F \le \frac{1}{20\kappa}$$

Proof. Since $\mathbf{M}$ is a rank k matrix, we know $\sigma_{k+1}(\mathbf{M}) = 0$, thus

$$\sigma_{k+1}(\frac{d_1 d_2}{m}\mathcal{P}_\Omega(\mathbf{M})) \le \sigma_{k+1}(\mathbf{M}) + \left\|\frac{d_1 d_2}{m}\mathcal{P}_\Omega(\mathbf{M}) - \mathbf{M}\right\| = \left\|\frac{d_1 d_2}{m}\mathcal{P}_\Omega(\mathbf{M}) - \mathbf{M}\right\|$$

Therefore:

$$\begin{aligned}
\left\|\mathbf{M} - \mathbf{U}_0\mathbf{V}_0^\top\right\| &\le \left\|\mathbf{M} - \frac{d_1 d_2}{m}\mathcal{P}_\Omega(\mathbf{M})\right\| + \left\|\frac{d_1 d_2}{m}\mathcal{P}_\Omega(\mathbf{M}) - \mathbf{U}_0\mathbf{V}_0^\top\right\| \\
&\le \left\|\mathbf{M} - \frac{d_1 d_2}{m}\mathcal{P}_\Omega(\mathbf{M})\right\| + \sigma_{k+1}(\frac{d_1 d_2}{m}\mathcal{P}_\Omega(\mathbf{M})) \le 2\left\|\mathbf{M} - \frac{d_1 d_2}{m}\mathcal{P}_\Omega(\mathbf{M})\right\|
\end{aligned}$$

Meanwhile, since $\text{rank}(\mathbf{M}) = k$, $\text{rank}(\mathbf{U}_0\mathbf{V}_0^\top) = k$, we know: $\text{rank}(\mathbf{M} - \mathbf{U}_0\mathbf{V}_0^\top) \le 2k$, and therefore:

$$\left\|\mathbf{M} - \mathbf{U}_0\mathbf{V}_0^\top\right\|_F \le \sqrt{2k}\left\|\mathbf{M} - \mathbf{U}_0\mathbf{V}_0^\top\right\| \le 2\sqrt{2k}\left\|\mathbf{M} - \frac{d_1 d_2}{m}\mathcal{P}_\Omega(\mathbf{M})\right\|$$

by choosing $m \ge c_0 \mu dk^2 \log d \cdot \kappa^2$ for large enough constant c_0 and apply Lemma A.3, we finishes the proof. □

A.2 Incoherence of Initialization

Lemma A.5. *Let $\mathbf{U}\mathbf{V}^\top$ be the top-k SVD of $\frac{d_1 d_2}{m}\mathcal{P}_\Omega(\mathbf{M})$, where $|\Omega| = m$. then there exists universal constant c_0, for any $m \geq c_0 \mu d k \kappa^2 \log d$, with probability at least $1 - \frac{1}{d^{10}}$, we have:*

$$\max_j \left\|\mathbf{e}_j^\top(\mathbf{M}^\top - \mathbf{V}\mathbf{U}^\top)\right\| \leq 2\sqrt{\frac{\mu k}{d_2}}$$

Proof. Suppose $\text{SVD}(\mathbf{M}) = \mathbf{X}\mathbf{S}\mathbf{Y}^\top$. Denote $\tilde{\mathbf{X}} = \mathbf{X}\mathbf{S}^{\frac{1}{2}}$ and $\tilde{\mathbf{Y}} = \mathbf{Y}\mathbf{S}^{\frac{1}{2}}$. Also let $\text{SVD}(\mathbf{U}\mathbf{V}^\top) = \mathbf{W}_\mathbf{U}\mathbf{D}\mathbf{W}_\mathbf{V}^\top$.

Then, we have:

$$\begin{aligned}
&\left\|\mathbf{e}_j^\top(\mathbf{M}^\top - \mathbf{V}\mathbf{U}^\top)\right\| = \left\|\mathbf{e}_j^\top(\mathbf{M}^\top - \frac{d_1 d_2}{m}\mathcal{P}_\Omega(\mathbf{M})^\top \mathbf{W}_\mathbf{U}\mathbf{W}_\mathbf{U}^\top)\right\| \\
=&\left\|\mathbf{e}_j^\top(\mathbf{M}^\top - \mathbf{M}^\top\mathbf{W}_\mathbf{U}\mathbf{W}_\mathbf{U}^\top + \mathbf{M}^\top\mathbf{W}_\mathbf{U}\mathbf{W}_\mathbf{U}^\top - \frac{d_1 d_2}{m}\mathcal{P}_\Omega(\mathbf{M})^\top \mathbf{W}_\mathbf{U}\mathbf{W}_\mathbf{U}^\top)\right\| \\
\leq&\left\|\mathbf{e}_j^\top\mathbf{M}^\top(\mathbf{I} - \mathbf{W}_\mathbf{U}\mathbf{W}_\mathbf{U}^\top)\right\| + \left\|\mathbf{e}_j^\top(\mathbf{M}^\top - \frac{d_1 d_2}{m}\mathcal{P}_\Omega(\mathbf{M})^\top)\mathbf{W}_\mathbf{U}\mathbf{W}_\mathbf{U}^\top\right\|
\end{aligned}$$

For the first term, since $\mathbf{W}_\mathbf{U}^\top\mathbf{W}_{\mathbf{U},\perp} = 0$, we have:

$$\begin{aligned}
&\left\|\mathbf{e}_j^\top\mathbf{M}^\top(\mathbf{I} - \mathbf{W}_\mathbf{U}\mathbf{W}_\mathbf{U}^\top)\right\| \leq \left\|\mathbf{e}_j^\top\mathbf{Y}\right\| \left\|\mathbf{S}\mathbf{X}^\top\mathbf{W}_{\mathbf{U},\perp}\mathbf{W}_{\mathbf{U},\perp}^\top\right\| \\
=&\sqrt{\frac{\mu k}{d_2}}\left\|\mathbf{Y}^\top(\mathbf{M}^\top - \mathbf{W}_\mathbf{V}\mathbf{D}\mathbf{W}_\mathbf{U}^\top)\mathbf{W}_{\mathbf{U},\perp}\mathbf{W}_{\mathbf{U},\perp}^\top\right\| \\
\leq&\sqrt{\frac{\mu k}{d_2}}\left\|\mathbf{M}^\top - \mathbf{W}_\mathbf{V}\mathbf{D}\mathbf{W}_\mathbf{U}^\top\right\| \leq \sqrt{\frac{\mu k}{d_2}}\cdot\frac{1}{\kappa}
\end{aligned}$$

The last step is due to sample $m \geq \mu d k \kappa^2 \log d$, and theorem A.4.

For the second term, we have:

$$\begin{aligned}
&\left\|\mathbf{e}_j^\top(\frac{d_1 d_2}{m}\mathcal{P}_\Omega(\mathbf{M})^\top - \mathbf{M}^\top)\mathbf{W}_\mathbf{U}\mathbf{W}_\mathbf{U}^\top\right\| = \left\|\tilde{\mathbf{Y}}_j^\top(\frac{d_1 d_2}{m}\sum_{i:(i,j)\in\Omega}\tilde{\mathbf{x}}_i\mathbf{w}_{\mathbf{U},i}^\top - \sum_i \tilde{\mathbf{x}}_i\mathbf{w}_{\mathbf{U},i}^\top)\mathbf{W}_\mathbf{U}^\top\right\| \\
\leq&\sqrt{\frac{\mu k}{d_2}}\cdot\frac{d_1 d_2}{m}\cdot\left\|\sum_{i:(i,j)\in\Omega}\tilde{\mathbf{x}}_i\mathbf{w}_{\mathbf{U},i}^\top - \frac{m}{d_1 d_2}\sum_i\tilde{\mathbf{x}}_i\mathbf{w}_{\mathbf{U},i}^\top\right\| \qquad (7)
\end{aligned}$$

Where $\tilde{\mathbf{x}}_i$ and $\mathbf{w}_{\mathbf{U},i}$ are the i-th row of $\tilde{\mathbf{X}}$ and $\mathbf{W}_\mathbf{U}$ respectively.

Let $\phi_{ij} = \tilde{\mathbf{x}}_i\mathbf{w}_{\mathbf{U},i}^\top(Z_{ij} - \frac{m}{d_1 d_2})$, where Z_{ij} is Bernoulli($\frac{m}{d_1 d_2}$) random variable, $Z_{ij} = 1$ iff $(i,j) \in \Omega$. Clearly, we have $\mathbb{E}\phi = 0$, and with probability 1:

$$\|\phi_{ij}\| \leq 2\left\|\tilde{\mathbf{x}}_i\right\|\left\|\mathbf{w}_{\mathbf{U},i}\right\| \leq 2\sqrt{\frac{\mu k}{d_1}}\max_i\left\|\mathbf{e}_i^\top\mathbf{W}_\mathbf{U}\right\|$$

Also, we have variance term:

$$\begin{aligned}
\left\|\sum_i\mathbb{E}\phi_{ij}^\top\phi_{ij}\right\| =& \left\|\sum_i\mathbb{E}(Z_{ij} - \frac{m}{d_1 d_2})^2\left\|\tilde{\mathbf{x}}_i\right\|^2\mathbf{w}_{\mathbf{U},i}\mathbf{w}_{\mathbf{U},i}^\top\right\| \\
\leq&\frac{m}{d_1 d_2}(1 - \frac{m}{d_1 d_2})\max_i\left\|\tilde{\mathbf{x}}_i\right\|^2\left\|\sum_i\mathbf{w}_{\mathbf{U},i}\mathbf{w}_{\mathbf{U},i}^\top\right\| \\
\leq&\frac{m}{d_1 d_2}\frac{\mu k}{d_1}\left\|\mathbf{W}_\mathbf{U}^\top\mathbf{W}_\mathbf{U}\right\| \leq \frac{\mu k m}{d_1^2 d_2} \\
\left\|\sum_i\mathbb{E}\phi_{ij}\phi_{ij}^\top\right\| =& \left\|\sum_i\mathbb{E}(Z_{ij} - \frac{m}{d_1 d_2})^2\left\|\mathbf{w}_{\mathbf{U},i}\right\|^2\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^\top\right\| \\
\leq&\frac{m}{d_1 d_2}\max_i\left\|\mathbf{e}_i^\top\mathbf{W}_\mathbf{U}\right\|^2
\end{aligned}$$

Therefore, with $m \geq \mu dk\kappa^2 \log d$, by matrix Bernstein, we have with probability at least $1 - \frac{1}{d^{10}}$, we know that for all $j \in [d_2]$, there exists some absolute constant C' so that:

$$\left\| \sum_{i:(i,j)\in\Omega} \tilde{\mathbf{x}}_i \mathbf{w}_{\mathbf{U},i}^\top - \frac{m}{d_1 d_2} \sum_i \tilde{\mathbf{x}}_i \mathbf{w}_{\mathbf{U},i}^\top \right\| \leq C' \sqrt{\frac{m \log d}{d_1 d_2}} (\sqrt{\frac{\mu k}{d_1}} + \max_i \left\| \mathbf{e}_i^\top \mathbf{W}_\mathbf{U} \right\|)$$

Substitue into Eq.(7), this gives:

$$\left\| \mathbf{e}_j^\top (\frac{d_1 d_2}{m} \mathcal{P}_\Omega(\mathbf{M})^\top - \mathbf{M}^\top) \mathbf{W}_\mathbf{U} \mathbf{W}_\mathbf{U}^\top \right\| \leq C' \sqrt{\frac{\mu k d_1 \log d}{m}} (\sqrt{\frac{\mu k}{d_1}} + \max_i \left\| \mathbf{e}_i^\top \mathbf{W}_\mathbf{U} \right\|)$$

On the other hand, we also have:

$$\begin{aligned} \left\| \mathbf{e}_i^\top \mathbf{W}_\mathbf{U} \right\| \leq & \left\| \mathbf{e}_i^\top \mathbf{W}_\mathbf{U} \mathbf{S} \right\| \left\| \mathbf{S}^{-1} \right\| = 2\kappa \left\| \mathbf{e}_i^\top \mathbf{U}\mathbf{V}^\top \right\| \leq 2\kappa (\left\| \mathbf{e}_i^\top (\mathbf{U}\mathbf{V}^\top - \mathbf{M}) \right\| + \left\| \mathbf{e}_i^\top \mathbf{M} \right\|) \\ \leq & 2\kappa (\sqrt{\frac{\mu k}{d_1}} + \left\| \mathbf{e}_i^\top (\mathbf{U}\mathbf{V}^\top - \mathbf{M}) \right\|) \end{aligned}$$

This gives overall inequality:

$$\max_j \left\| \mathbf{e}_j^\top (\mathbf{V}\mathbf{U}^\top - \mathbf{M}^\top) \right\| \leq \sqrt{\frac{\mu k}{d_2}} \cdot \frac{1}{\kappa} + C'' \sqrt{\frac{\mu k d_1 \log d}{m}} \kappa (\sqrt{\frac{\mu k}{d_1}} + \max_i \left\| \mathbf{e}_i^\top (\mathbf{U}\mathbf{V}^\top - \mathbf{M}) \right\|)$$

By symmetry, we will also have:

$$\max_i \left\| \mathbf{e}_i^\top (\mathbf{U}\mathbf{V}^\top - \mathbf{M}) \right\| \leq \sqrt{\frac{\mu k}{d_1}} \cdot \frac{1}{\kappa} + C'' \sqrt{\frac{\mu k d_2 \log d}{m}} \kappa (\sqrt{\frac{\mu k}{d_2}} + \max_j \left\| \mathbf{e}_j^\top (\mathbf{V}\mathbf{U}^\top - \mathbf{M}^\top) \right\|)$$

Combine above two equations and choose $m \geq c_0 \mu dk\kappa^2 \log d$ for some large enough c_0. We have:

$$\max_j \left\| \mathbf{e}_j^\top (\mathbf{M}^\top - \mathbf{V}\mathbf{U}^\top) \right\| \leq 2\sqrt{\frac{\mu k}{d_2}}$$

This finishes the proof.

□

Theorem A.6. *Let $\mathbf{U}_0 \mathbf{V}_0^\top$ be the top-k SVD of $\frac{d_1 d_2}{m} \mathcal{P}_\Omega(\mathbf{M})$, where $|\Omega| = m$. then there exists universal constant c_0, for any $m \geq c_0 \mu dk\kappa^2 \log d$, with probability at least $1 - \frac{1}{d^{10}}$, we have:*

$$\max_i \left\| \mathbf{e}_i^\top \mathbf{U}_0 \mathbf{V}_0^\top \right\|^2 \leq \frac{9\mu k}{d_1} \quad \textit{and} \quad \max_j \left\| \mathbf{e}_j^\top \mathbf{V}_0 \mathbf{U}_0^\top \right\|^2 \leq \frac{9\mu k}{d_2}$$

Proof. By Theorem A.5, we know for any $j \in [d_2]$:

$$\left\| \mathbf{e}_j^\top (\mathbf{M}^\top - \mathbf{V}_0 \mathbf{U}_0^\top) \right\| \leq 2\sqrt{\frac{\mu k}{d_2}}$$

Therefore, we have:

$$\left\| \mathbf{e}_j^\top \mathbf{V}_0 \mathbf{U}_0^\top \right\| \leq [\left\| \mathbf{e}_j^\top \mathbf{M}^\top \right\| + \left\| \mathbf{e}_j^\top (\mathbf{M}^\top - \mathbf{V}_0 \mathbf{U}_0^\top) \right\|] \leq 3\sqrt{\frac{\mu k}{d_2}}$$

By symmetry, we also know for any $i \in [d_1]$

$$\left\| \mathbf{e}_i^\top \mathbf{U}_0 \mathbf{V}_0^\top \right\| \leq 3\sqrt{\frac{\mu k}{d_1}}$$

Which finishes the proof. □

For the special case where $\mathbf{M} \in \mathbb{R}^{d \times d}$ is symmetric and PSD, we can easily extends to have following:

Corollary A.7. *Let $\mathbf{U}_0\mathbf{U}_0^\top$ be the top-k SVD of $\frac{d^2}{m}\mathcal{P}_\Omega(\mathbf{M})$, where $|\Omega| = m$. then there exists universal constant c_0, for any $m \geq c_0\mu dk\kappa^2 \log d$, with probability at least $1 - \frac{1}{d^{10}}$, we have:*

$$\max_i \left\|\mathbf{e}_i^\top \mathbf{U}_0\right\|^2 \leq \frac{10\mu k\kappa}{d}$$

Proof. By Corollary A.6, we have:

$$\max_i \left\|\mathbf{e}_i^\top \mathbf{U}_0\mathbf{U}_0^\top\right\|^2 \leq \frac{9\mu k}{d}$$

On the other hand, by Theorem A.4, we have:

$$\sigma_{\min}(\mathbf{U}_0^\top\mathbf{U}_0) = \sigma_k(\mathbf{U}_0\mathbf{U}_0^\top) \geq \sigma_k(\mathbf{M}) - \left\|\mathbf{M} - \mathbf{U}_0\mathbf{U}_0^\top\right\| \geq \frac{9}{10\kappa}$$

Therefore, for any $i \in [d]$ we have:

$$\left\|\mathbf{e}_i^\top \mathbf{U}_0\right\|^2 \leq \frac{\left\|\mathbf{e}_i^\top \mathbf{U}_0\mathbf{U}_0^\top\right\|^2}{\sigma_{\min}(\mathbf{U}_0^\top\mathbf{U}_0)} \leq \frac{10\mu k\kappa}{d}$$

Which finishes the proof. □

Finally, Lemma A.1 can be easily concluded from Theorem A.4 and Theorem A.6, while Lemma 4.1 is also directly proved by Theorem A.4 and Corollary A.7.

B Proof of Symmetric PSD Case

In this section, we prove Theorem 3.1. WLOG, we continue to assume $\|\mathbf{M}\| = 1$ in all proof. Also, when it's clear from the context, we use κ to specifically to represent $\kappa(\mathbf{M})$. Then $\sigma_{\min}(\mathbf{M}) = \frac{1}{\kappa}$. Also in this section, we always denote SVD$(\mathbf{M}) = \mathbf{X}\mathbf{S}\mathbf{X}^\top$, and SVD$(\mathbf{U}\mathbf{U}^\top) = \mathbf{W}\mathbf{D}\mathbf{W}^\top$.

The most essential part to prove Theorem 3.1 is proving following Theorem:

Theorem B.1 (restatement of Theorem 4.5)**.** *Let $f(\mathbf{U}) = \left\|\mathbf{U}\mathbf{U}^\top - \mathbf{M}\right\|_F^2$ and $g_i(\mathbf{U}) = \left\|\mathbf{e}_i^\top\mathbf{U}\right\|^2$. Suppose after initialization, we have:*

$$f(\mathbf{U}_0) \leq \left(\frac{1}{20\kappa}\right)^2, \qquad \max_i g_i(\mathbf{U}_0) \leq \frac{10\mu k\kappa^2}{d}$$

Then, there exist some absolute constant c such that for any learning rate $\eta < \frac{c}{\mu dk\kappa^3 \log d}$, with at least $1 - \frac{T}{d^{10}}$ probability, we will have for all $t \leq T$ that:

$$f(\mathbf{U}_t) \leq (1 - \frac{\eta}{2\kappa})^t \left(\frac{1}{10\kappa}\right)^2, \qquad \max_i g_i(\mathbf{U}_t) \leq \frac{20\mu k\kappa^2}{d}$$

Theorem B.1 says once initialization algorithm provides $\mathbf{U}_0$ in good local region, with high probability $\mathbf{U}_t$ will always stay in this good region and $f(\mathbf{U}_t)$ is linear converging to 0. With this theorem, we can then immediately conclude Theorem 3.1 from Theorem B.1 and Lemma 4.1.

The rest of this section all focus on proving Theorem B.1. First, we prepare with a few lemmas about the property of objective function, and the spectral property of $\mathbf{U}$ in a local Frobenius ball around optimal. Then, we prove Theorem B.1 by constructing two supermartingales related to $f(\mathbf{U}_t), g_i(\mathbf{U}_t)$ each, and applying concentration argument.

For symmetric PSD case, we denote the stochastic gradient as:

$$SG(\mathbf{U}) = 2d^2(\mathbf{U}\mathbf{U}^\top - \mathbf{M})_{ij}(\mathbf{e}_i\mathbf{e}_j^\top + \mathbf{e}_j\mathbf{e}_i^\top)\mathbf{U}$$

The update in Algorithm 1 can be now written as:

$$\mathbf{U}_{t+1} \leftarrow \mathbf{U}_t - \eta SG(\mathbf{U}_t) \tag{8}$$

We immediately have the property:

$$\mathbb{E}SG(\mathbf{U}) = \nabla f(\mathbf{U}) = 4(\mathbf{U}\mathbf{U}^\top - \mathbf{M})\mathbf{U}$$

B.1 Geometric Properties in Local Region

First, we prove two lemmas w.r.t the smoothness and property similar to strongly convex for objective function:

Lemma B.2. *(restatement of Lemma 4.2) Within the region $\mathcal{D} = \{\mathbf{U} | \|\mathbf{U}\| \leq \Gamma\}$, we have function $f(\mathbf{U}) = \|\mathbf{M} - \mathbf{U}\mathbf{U}^\top\|_F^2$ satisfying for any $\mathbf{U}_1, \mathbf{U}_2 \in \mathcal{D}$:*

$$\|\nabla f(\mathbf{U}_1) - \nabla f(\mathbf{U}_2)\|_F \leq \beta \|\mathbf{U}_1 - \mathbf{U}_2\|_F$$

where smoothness parameter $\beta = 16 \max\{\Gamma^2, \|\mathbf{M}\|\}$.

Proof. Inside region $\mathcal{D}$, we have:

$$\begin{aligned}
&\|\nabla f(\mathbf{U}_1) - \nabla f(\mathbf{U}_2)\|_{\mathrm{F}} \\
=&\|4(\mathbf{U}_1\mathbf{U}_1^\top - \mathbf{M})\mathbf{U}_1 - 4(\mathbf{U}_2\mathbf{U}_2^\top - \mathbf{M})\mathbf{U}_2\|_{\mathrm{F}} \\
\leq&4\|\mathbf{U}_1\mathbf{U}_1^\top\mathbf{U}_1 - \mathbf{U}_2\mathbf{U}_2^\top\mathbf{U}_2\|_{\mathrm{F}} + 4\|\mathbf{M}(\mathbf{U}_1 - \mathbf{U}_2)\|_{\mathrm{F}} \\
=&4\|\mathbf{U}_1\mathbf{U}_1^\top(\mathbf{U}_1 - \mathbf{U}_2) + \mathbf{U}_1(\mathbf{U}_1 - \mathbf{U}_2)^\top\mathbf{U}_2 + (\mathbf{U}_1 - \mathbf{U}_2)\mathbf{U}_2^\top\mathbf{U}_2\|_{\mathrm{F}} + 4\|\mathbf{M}(\mathbf{U}_1 - \mathbf{U}_2)\|_{\mathrm{F}} \\
\leq&12\max\{\|\mathbf{U}_1\|^2, \|\mathbf{U}_2\|^2\}\|\mathbf{U}_1 - \mathbf{U}_2\|_{\mathrm{F}} + 4\|\mathbf{M}\|\,\|\mathbf{U}_1 - \mathbf{U}_2\|_{\mathrm{F}} \\
\leq&16\max\{\Gamma^2, \|\mathbf{M}\|\}\|\mathbf{U}_1 - \mathbf{U}_2\|_{\mathrm{F}}
\end{aligned}$$

□

Lemma B.3. *(restatement of Lemma 4.3) Within the region $\mathcal{D} = \{\mathbf{U} | \sigma_{\min}(\mathbf{X}^\top\mathbf{U}) \geq \gamma\}$, then we have function $f(\mathbf{U}) = \|\mathbf{M} - \mathbf{U}\mathbf{U}^\top\|_F^2$ satisfying:*

$$\|\nabla f(\mathbf{U})\|_F^2 \geq \alpha f(\mathbf{U})$$

where constant $\alpha = 4\gamma^2$.

Proof. Inside region $\mathcal{D}$, recall we denote $\mathbf{W}\mathbf{D}\mathbf{W}^\top = \mathrm{SVD}(\mathbf{U}\mathbf{U}^\top)$, thus we have:

$$\begin{aligned}
&\|\nabla f(\mathbf{U})\|_{\mathrm{F}}^2 = 16\|(\mathbf{U}\mathbf{U}^\top - \mathbf{M})\mathbf{U}\|_{\mathrm{F}}^2 \\
=&16[\|\mathcal{P}_{\mathbf{W}}(\mathbf{U}\mathbf{U}^\top - \mathbf{M})\mathbf{U}\|_{\mathrm{F}}^2 + \|\mathcal{P}_{\mathbf{W}_\perp}(\mathbf{U}\mathbf{U}^\top - \mathbf{M})\mathbf{U}\|_{\mathrm{F}}^2] \\
\geq&16[\sigma_{\min}(\mathbf{D})\|\mathcal{P}_{\mathbf{W}}(\mathbf{U}\mathbf{U}^\top - \mathbf{M})\mathcal{P}_{\mathbf{W}}\|_{\mathrm{F}}^2 + \|\mathcal{P}_{\mathbf{W}_\perp}\mathbf{M}\mathbf{U}\|_{\mathrm{F}}^2] \\
\geq&16[\sigma_{\min}(\mathbf{D})\|\mathbf{U}\mathbf{U}^\top - \mathcal{P}_{\mathbf{W}}\mathbf{M}\mathcal{P}_{\mathbf{W}}\|_{\mathrm{F}}^2 + \|\mathcal{P}_{\mathbf{W}_\perp}\mathbf{M}\mathbf{U}\|_{\mathrm{F}}^2]
\end{aligned}$$

On the other hand, we have:

$$\begin{aligned}
&\|\mathcal{P}_{\mathbf{W}_\perp}\mathbf{M}\mathbf{U}\|_{\mathrm{F}}^2 = \|\mathcal{P}_{\mathbf{W}_\perp}\mathbf{X}\mathbf{\Sigma}\mathbf{X}^\top\mathbf{U}\|_{\mathrm{F}}^2 \geq \sigma_{\min}^2(\mathbf{X}^\top\mathbf{U})\|\mathcal{P}_{\mathbf{W}_\perp}\mathbf{X}\mathbf{\Sigma}\|_{\mathrm{F}}^2 \\
=&\sigma_{\min}^2(\mathbf{X}^\top\mathbf{U})\mathrm{tr}(\mathcal{P}_{\mathbf{W}_\perp}\mathbf{M}^2\mathcal{P}_{\mathbf{W}_\perp}) = \sigma_{\min}^2(\mathbf{X}^\top\mathbf{U})\|\mathcal{P}_{\mathbf{W}_\perp}\mathbf{M}\|_{\mathrm{F}}^2
\end{aligned}$$

and

$$\sigma_{\min}(\mathbf{D}) = \lambda_{\min}(\mathbf{U}^\top\mathbf{U}) \geq \lambda_{\min}(\mathbf{U}^\top\mathcal{P}_{\mathbf{X}}\mathbf{U}) = \sigma_{\min}^2(\mathbf{X}^\top\mathbf{U})$$

Therefore, combine all above, we have:

$$\begin{aligned}
&\|\nabla f(\mathbf{U})\|_{\mathrm{F}}^2 \geq 16\sigma_{\min}^2(\mathbf{X}^\top\mathbf{U})[\|\mathbf{U}\mathbf{U}^\top - \mathcal{P}_{\mathbf{W}}\mathbf{M}\mathcal{P}_{\mathbf{W}}\|_{\mathrm{F}}^2 + \|\mathcal{P}_{\mathbf{W}_\perp}\mathbf{M}\|_{\mathrm{F}}^2] \\
\geq&4\sigma_{\min}^2(\mathbf{X}^\top\mathbf{U})[\|\mathbf{U}\mathbf{U}^\top - \mathcal{P}_{\mathbf{W}}\mathbf{M}\mathcal{P}_{\mathbf{W}}\|_{\mathrm{F}}^2 + \|\mathcal{P}_{\mathbf{W}_\perp}\mathbf{M}\mathcal{P}_{\mathbf{W}}\|_{\mathrm{F}}^2 + \|\mathcal{P}_{\mathbf{W}}\mathbf{M}\mathcal{P}_{\mathbf{W}_\perp}\|_{\mathrm{F}}^2 + \|\mathcal{P}_{\mathbf{W}_\perp}\mathbf{M}\mathcal{P}_{\mathbf{W}_\perp}\|_{\mathrm{F}}^2] \\
=&4\sigma_{\min}^2(\mathbf{X}^\top\mathbf{U})\|\mathbf{U}\mathbf{U}^\top - \mathbf{M}\|_{\mathrm{F}}^2
\end{aligned}$$

□

Next, we show as long as we are in some Frobenious ball around optimum, then we have good spectral property over $\mathbf{U}$ which guarantees the preconditions for Lemma B.2 and Lemma B.3.

Lemma B.4. *(restatement of Lemma 4.4) Within the region $\mathcal{D} = \{\mathbf{U} | \left\|\mathbf{M} - \mathbf{U}\mathbf{U}^\top\right\|_F \leq \frac{1}{10}\sigma_k(\mathbf{M})\}$, we have:*

$$\|\mathbf{U}\| \leq \sqrt{2\|\mathbf{M}\|}, \qquad \sigma_{\min}(\mathbf{X}^\top\mathbf{U}) \geq \sqrt{\sigma_k(\mathbf{M})/2}$$

Proof. For spectral norm of $\mathbf{U}$, we have:

$$\|\mathbf{U}\|^2 \leq \|\mathbf{M}\| + \left\|\mathbf{M} - \mathbf{U}\mathbf{U}^\top\right\| \leq \|\mathbf{M}\| + \left\|\mathbf{M} - \mathbf{U}\mathbf{U}^\top\right\|_F \leq 2\|\mathbf{M}\|$$

For the minimum singular value of $\mathbf{U}^\top\mathbf{U}$, we have:

$$\begin{aligned}\sigma_{\min}(\mathbf{U}^\top\mathbf{U}) =& \sigma_k(\mathbf{U}\mathbf{U}^\top) \geq \sigma_k(\mathbf{M}) - \left\|\mathbf{M} - \mathbf{U}\mathbf{U}^\top\right\| \\ \geq& \sigma_k(\mathbf{U}\mathbf{U}^\top) \geq \sigma_k(\mathbf{M}) - \left\|\mathbf{M} - \mathbf{U}\mathbf{U}^\top\right\|_F \geq \frac{9}{10}\sigma_k(\mathbf{M})\end{aligned}$$

On the other hand, we have:

$$\begin{aligned}\frac{9}{10}\sigma_k(\mathbf{M})\|\mathbf{X}_\perp\mathbf{W}\|^2 \leq& \sigma_{\min}(\mathbf{D})\|\mathbf{X}_\perp\mathbf{W}\|^2 \leq \left\|\mathbf{X}_\perp^\top\mathbf{W}\mathbf{\Sigma}\mathbf{W}^\top\mathbf{X}_\perp\right\| \\ \leq& \left\|\mathbf{X}_\perp^\top\mathbf{U}\mathbf{U}^\top\mathbf{X}_\perp\right\|_F = \left\|\mathcal{P}_{\mathbf{X}_\perp}(\mathbf{M} - \mathbf{U}\mathbf{U}^\top)\mathcal{P}_{\mathbf{X}_\perp}\right\|_F \\ \leq& \left\|\mathbf{M} - \mathbf{U}\mathbf{U}^\top\right\|_F \leq \frac{1}{10}\sigma_k(\mathbf{M})\end{aligned}$$

Let the principal angle between $\mathbf{X}$ and $\mathbf{W}$ to be θ. This gives $\sin^2\theta = \left\|\mathbf{X}_\perp^\top\mathbf{W}\right\|^2 \leq \frac{1}{9}$. Thus $\cos^2\theta = \sigma_{\min}^2(\mathbf{X}^\top\mathbf{W}) \geq \frac{8}{9}$. Therefore:

$$\sigma_{\min}^2(\mathbf{X}^\top\mathbf{U}) \geq \sigma_{\min}^2(\mathbf{X}^\top\mathbf{W})\sigma_{\min}(\mathbf{U}^\top\mathbf{U}) \geq \sigma_k(\mathbf{M})/2$$

□

B.2 Proof of Theorem B.1

Now, we are ready for our key theorem. By Lemma B.2, Lemma B.3, and Lemma B.4, we already know the function has good property locally in the region $\mathcal{D} = \{\mathbf{U} | \left\|\mathbf{M} - \mathbf{U}\mathbf{U}^\top\right\|_F \leq \frac{1}{10}\sigma_k(\mathbf{M})\}$ which alludes linear convergence. Then, the work remains and also the most challenging part is to prove that once we initialize inside this region, our algorithm will guarantee $\mathbf{U}$ never leave this region with high probability even with relatively large stepsize. The requirement for tight sample complexity and near optimal runtime makes it more challenging, and require us to further control the incoherence of $\mathbf{U}_t$ over all iterates in addition to the distance $\left\|\mathbf{M} - \mathbf{U}\mathbf{U}^\top\right\|_F$.

Following is our formal proof.

Proof of Theorem B.1. Define event $\mathfrak{E}_t = \{\forall\tau \leq t, f(\mathbf{U}_\tau) \leq (1 - \frac{\eta}{2\kappa})^t(\frac{1}{10\kappa})^2, \max_i g_i(\mathbf{U}_\tau) \leq \frac{20\mu k\kappa^2}{d}\}$. Theorem B.1 is equivalent to prove event $\mathfrak{E}_T$ happens with high probability. The proof achieves this by contructing two supermartingales for $f(\mathbf{U}_t)1_{\mathfrak{E}_t}$ and $g_i(\mathbf{U}_t)1_{\mathfrak{E}_t}$ (where $1_{(\cdot)}$ denote indicator function), applies concentration argument.

The proofs follow the structure of:

1. The constructions of supermartingales
2. Their probability 1 bound and variance bound in order to apply Azuma-Bernstein inequality
3. Final combination of concentration results to conclude the proof

First, let filtration $\mathfrak{F}_t = \sigma\{SG(\mathbf{U}_0), \cdots, SG(\mathbf{U}_{t-1})\}$ where $\sigma\{\cdot\}$ denotes the sigma field. Note by definiton of $\mathfrak{E}_t$, we have $\mathfrak{E}_t \subset \mathfrak{F}_t$. Also $\mathfrak{E}_{t+1} \subset \mathfrak{E}_t$, and thus $1_{\mathfrak{E}_{t+1}} \leq 1_{\mathfrak{E}_t}$. Note $\mathfrak{E}_t$ denotes the event which up to time t, $\mathbf{U}_\tau$ always stay in a local region which both close to $\mathbf{M}$ and incoherent.

By Lemma B.4, we immediately know that conditioned on $\mathfrak{E}_t$, we have $\|\mathbf{U}_t\| \leq \sqrt{2}$, $\sigma_{\min}(\mathbf{X}^\top\mathbf{U}_t) \geq 1/\sqrt{2\kappa}$ and $\sigma_{\min}(\mathbf{U}_t^\top\mathbf{U}_t) \geq 1/2\kappa$. We will use this fact throughout the proof.

Construction of supermartingale G: Since $g_i(\mathbf{U}) = \mathbf{e}_i^\top \mathbf{U}\mathbf{U}^\top \mathbf{e}_i$ is a quadratic function, we know for any change $\Delta\mathbf{U}$, we have:

$$g_i(\mathbf{U}+\Delta\mathbf{U}) = g_i(\mathbf{U}) + 2\mathbf{e}_i^\top(\Delta\mathbf{U})\mathbf{U}^\top\mathbf{e}_i + \left\|\mathbf{e}_i^\top\Delta\mathbf{U}\right\|^2$$

We know for any $l \in [d]$:

$$\begin{aligned}&\mathbb{E}\|\mathbf{e}_l^\top SG(\mathbf{U})\|^2 1_{\mathfrak{E}_t} \le \mathbb{E}16d^4\delta_{il}(\mathbf{u}_i^\top\mathbf{u}_j - \mathbf{M}_{ij})^2 \max_i \left\|\mathbf{e}_i^\top\mathbf{U}\right\|^2 1_{\mathfrak{E}_t}\\ =&16d^2\left\|\mathbf{e}_l^\top(\mathbf{U}\mathbf{U}^\top-\mathbf{M})\right\|^2 \max_i\left\|\mathbf{e}_i^\top\mathbf{U}\right\|^2 1_{\mathfrak{E}_t} \le O(\mu^2k^2\kappa^4)1_{\mathfrak{E}_t}\end{aligned}$$

Therefore, by update Eq.(8), and $\mathbb{E}SG(\mathbf{U}) = \nabla f(\mathbf{U}) = 4(\mathbf{U}\mathbf{U}^\top-\mathbf{M})\mathbf{U}$, we know:

$$\begin{aligned}&\mathbb{E}[g_i(\mathbf{U}_{t+1})1_{\mathfrak{E}_t}|\mathfrak{F}_t]\\ =&[g_i(\mathbf{U}_t) - 2\eta\mathbf{e}_i^\top[\mathbb{E}SG(\mathbf{U}_t)]\mathbf{U}_t^\top\mathbf{e}_i + \frac{\eta^2}{2}\mathbb{E}\left\|\mathbf{e}_i^\top SG(\mathbf{U}_t)\right\|^2]1_{\mathfrak{E}_t}\\ =&[\text{tr}(\mathbf{U}_t^\top\mathbf{e}_i\mathbf{e}_i^\top[\mathbf{I}-8\eta(\mathbf{U}_t\mathbf{U}_t^\top-\mathbf{M})]\mathbf{U}_t) + \frac{\eta^2}{2}\mathbb{E}\left\|\mathbf{e}_i^\top SG(\mathbf{U}_t)\right\|^2]1_{\mathfrak{E}_t}\\ =&[\text{tr}(\mathbf{U}_t^\top\mathbf{e}_i\mathbf{e}_i^\top\mathbf{U}_t(\mathbf{I}-8\eta\mathbf{U}_t^\top\mathbf{U}_t)) + 8\eta\text{tr}(\mathbf{U}_t^\top\mathbf{e}_i\mathbf{e}_i^\top\mathbf{M}\mathbf{U}_t) + \eta^2O(\mu^2k^2\kappa^4)]1_{\mathfrak{E}_t}\\ \le&[(1-8\eta\sigma_{\min}(\mathbf{U}_t^\top\mathbf{U}_t))g_i(\mathbf{U}_t) + 8\eta\text{tr}(\mathbf{U}_t^\top\mathbf{e}_i\mathbf{e}_i^\top\mathbf{M}\mathbf{U}_t) + \eta^2O(\mu^2k^2\kappa^4)]1_{\mathfrak{E}_t}\\ \le&[(1-\frac{4\eta}{\kappa})g_i(\mathbf{U}_t) + 16\sqrt{10}\frac{\eta\mu k\kappa}{d} + \eta^2O(\mu^2k^2\kappa^4)]1_{\mathfrak{E}_t}\\ \le&[(1-\frac{4\eta}{\kappa})g_i(\mathbf{U}_t) + 60\frac{\eta\mu k\kappa}{d})]1_{\mathfrak{E}_t}\end{aligned}$$

The last step is true by choosing constant c in learning rate η to be small enough.

Let $G_{it} = (1-\frac{4\eta}{\kappa})^{-t}(g_i(\mathbf{U}_t)1_{\mathfrak{E}_{t-1}} - 15\frac{\mu k\kappa^2}{d})$. This gives:

$$\mathbb{E}G_{i(t+1)} \le (1-\frac{4\eta}{\kappa})^{-t}(g_i(\mathbf{U}_t)1_{\mathfrak{E}_t} - 15\frac{\mu k\kappa^2}{d}) \le G_{it}$$

That is G_{it} is supermartingale.

Probability 1 bound for G: We also know

$$\begin{aligned}G_{it} - \mathbb{E}[G_{it}|\mathfrak{F}_{t-1}] =&(1-\frac{4\eta}{\kappa})^{-t}\Big[-\eta\mathbf{e}_i^\top[SG(\mathbf{U}_t)-\mathbb{E}SG(\mathbf{U}_t)]\mathbf{U}_t^\top\mathbf{e}_i\\ &+\frac{\eta^2}{2}[\|\mathbf{e}_i^\top SG(\mathbf{U}_t)\|^2 - \mathbb{E}\|\mathbf{e}_i^\top SG(\mathbf{U}_t)\|^2]\Big]1_{\mathfrak{E}_{t-1}}\end{aligned} \tag{9}$$

Since when sample (i,j) entry of matrix $\mathbf{M}$, for any $l \in [d]$, we have:

$$\begin{aligned}&\mathbf{e}_l^\top[SG(\mathbf{U}_t)]\mathbf{U}_t^\top\mathbf{e}_l\cdot 1_{\mathfrak{E}_{t-1}} = O(1)\text{tr}(\mathbf{U}^\top\mathbf{e}_l\mathbf{e}_l^\top SG(\mathbf{U}_t))1_{\mathfrak{E}_{t-1}}\\ =&O(1)d^2(\mathbf{U}\mathbf{U}^\top-\mathbf{M})_{ij}\text{tr}[\mathbf{U}^\top\mathbf{e}_l\mathbf{e}_l^\top(\mathbf{e}_i\mathbf{u}_j^\top+\mathbf{e}_j\mathbf{u}_i^\top)]1_{\mathfrak{E}_{t-1}}\\ \le&O(1)d^2\left\|\mathbf{U}\mathbf{U}^\top-\mathbf{M}\right\|_\infty\max_i\left\|\mathbf{e}_i^\top\mathbf{U}\right\|^2 1_{\mathfrak{E}_{t-1}} \le O(\mu^2k^2\kappa^4)1_{\mathfrak{E}_{t-1}}\end{aligned}$$

and

$$\begin{aligned}&\|\mathbf{e}_l^\top SG(\mathbf{U}_t)\|^2 1_{\mathfrak{E}_{t-1}} = O(1)\left\|\mathbf{e}_l^\top SG(\mathbf{U}_t)\right\|^2 1_{\mathfrak{E}_{t-1}}\\ =&O(1)d^4(\mathbf{U}\mathbf{U}^\top-\mathbf{M})_{ij}^2\left\|\mathbf{e}_l^\top(\mathbf{e}_i\mathbf{u}_j^\top+\mathbf{e}_j\mathbf{u}_i^\top)\right\|^2 1_{\mathfrak{E}_{t-1}}\\ \le&O(1)d^4\left\|\mathbf{U}\mathbf{U}^\top-\mathbf{M}\right\|_\infty^2\max_i\left\|\mathbf{e}_i^\top\mathbf{U}\right\|^2 1_{\mathfrak{E}_{t-1}} \le O(\mu^3dk^3\kappa^6)1_{\mathfrak{E}_{t-1}}\end{aligned}$$

Therefore, by Eq.(9), we have with probability 1:

$$|G_{it} - \mathbb{E}[G_{it}|\mathfrak{F}_{t-1}]| \le (1-\frac{4\eta}{\kappa})^{-t}\eta O(\mu^2k^2\kappa^4)1_{\mathfrak{E}_{t-1}} \tag{10}$$

Variance bound for G: For any $l \in [d]$, we also know

$$\begin{aligned}
&\text{Var}(\mathbf{e}_l^\top [SG(\mathbf{U}_t)]\mathbf{U}_t^\top \mathbf{e}_l \cdot 1_{\mathfrak{E}_{t-1}} | \mathfrak{F}_{t-1}) \leq \mathbb{E}[\langle \nabla g_l(\mathbf{U}_t), SG(\mathbf{U}_t)\rangle^2 1_{\mathfrak{E}_{t-1}} | \mathfrak{F}_{t-1}] \\
=&O(1)\frac{1}{d^2}\sum_{ij} d^4 (\mathbf{U}\mathbf{U}^\top - \mathbf{M})_{ij}^2 \text{tr}[\mathbf{U}^\top \mathbf{e}_l \mathbf{e}_l^\top (\mathbf{e}_i \mathbf{u}_j^\top + \mathbf{e}_j \mathbf{u}_i^\top)]^2 1_{\mathfrak{E}_{t-1}} \\
\leq& O(1) d^2 \sum_j (\mathbf{U}\mathbf{U}^\top - \mathbf{M})_{lj}^2 \text{tr}[\mathbf{U}^\top \mathbf{e}_l \mathbf{u}_j^\top]^2 1_{\mathfrak{E}_{t-1}} \\
\leq& O(1) d^2 \left\|\mathbf{e}_l^\top (\mathbf{U}\mathbf{U}^\top - \mathbf{M})\right\|^2 \max_i \left\|\mathbf{e}_i^\top \mathbf{U}\right\|^4 1_{\mathfrak{E}_{t-1}} \leq O(\frac{\mu^3 k^3 \kappa^6}{d}) 1_{\mathfrak{E}_{t-1}}
\end{aligned}$$

and

$$\begin{aligned}
&\text{Var}(\|\mathbf{e}_l^\top SG(\mathbf{U})\|^2 1_{\mathfrak{E}_{t-1}} | \mathfrak{F}_{t-1}) \leq \mathbb{E}[\nabla^2 g_k(SG(\mathbf{U}_t), SG(\mathbf{U}_t))^2 1_{\mathfrak{E}_{t-1}} | \mathfrak{F}_{t-1}] \\
=&O(1)\frac{1}{d^2}\sum_{ij} d^8 (\mathbf{U}\mathbf{U}^\top - \mathbf{M})_{ij}^4 \left\|\mathbf{e}_k^\top (\mathbf{e}_i \mathbf{u}_j^\top + \mathbf{e}_j \mathbf{u}_i^\top)\right\|^4 1_{\mathfrak{E}_{t-1}} \\
\leq& O(1) d^6 \sum_j (\mathbf{U}\mathbf{U}^\top - \mathbf{M})_{kj}^4 \left\|\mathbf{u}_j\right\|^4 1_{\mathfrak{E}_{t-1}} \\
\leq& O(1) d^6 \left\|\mathbf{U}\mathbf{U}^\top - \mathbf{M}\right\|_\infty^2 \left\|\mathbf{e}_k^\top (\mathbf{U}\mathbf{U}^\top - \mathbf{M})\right\|^2 \max_i \left\|\mathbf{e}_i^\top \mathbf{U}\right\|^4 1_{\mathfrak{E}_{t-1}} \leq O(\mu^5 d k^5 \kappa^{10}) 1_{\mathfrak{E}_{t-1}}
\end{aligned}$$

Therefore, by Eq.(9), we have

$$\text{Var}(G_{it} | \mathfrak{F}_{t-1}) \leq (1 - \frac{4\eta}{\kappa})^{-2t} \eta^2 O(\frac{\mu^3 k^3 \kappa^6}{d}) 1_{\mathfrak{E}_{t-1}} \tag{11}$$

Bernstein's inequality for G: Let $\sigma^2 = \sum_{\tau=1}^t \text{Var}(G_{i\tau} | \mathfrak{F}_{\tau-1})$, and R satisfies, with probability 1 that $|G_{i\tau} - \mathbb{E}[G_{i\tau} | \mathfrak{F}_{\tau-1}]| \leq R,\ \tau = 1, \cdots, t$. Then By standard Bernstein concentration inequality, we know:

$$P(G_{it} \geq G_{i0} + s) \leq \exp(\frac{s^2/2}{\sigma^2 + Rs/3})$$

Since $G_{i0} = g_i(\mathbf{U}_0) - 15\frac{\mu k \kappa^2}{d}$, let $s' = O(1)(1 - \frac{4\eta}{\kappa})^t [\sqrt{\sigma^2 \log d} + R \log d]$, we know

$$P\left(g_i(\mathbf{U}_t) 1_{\mathfrak{E}_{t-1}} \geq 15\frac{\mu k \kappa^2}{d} + (1 - \frac{4\eta}{\kappa})^t (g_i(\mathbf{U}_0) - 15\frac{\mu k \kappa^2}{d}) + s'\right) \leq \frac{1}{2d^{11}}$$

By Eq.(10), we know $R = (1 - \frac{4\eta}{\kappa})^{-t} \eta O(\mu^2 k^2 \kappa^4)$ satisfies that $|G_{i\tau} - \mathbb{E}[G_{i\tau} | \mathfrak{F}_{\tau-1}]| \leq R,\ \tau = 1, \cdots, t$. Also by Eq. (11), we have:

$$(1 - \frac{4\eta}{\kappa})^t \sqrt{\sigma^2 \log d} \leq \eta O(\sqrt{\frac{\mu^3 k^3 \kappa^6 \log d}{d}}) \sqrt{\sum_{\tau=1}^t (1 - \frac{4\eta}{\kappa})^{2t-2\tau}} \leq \sqrt{\eta} O(\sqrt{\frac{\mu^3 k^3 \kappa^7 \log d}{d}})$$

by $\eta < \frac{c}{\mu d k \kappa^3 \log d}$ and choosing c to be small enough, we have:

$$s' = \sqrt{\eta} O(\sqrt{\frac{\mu^3 k^3 \kappa^7 \log d}{d}}) + \eta O(\mu^2 k^2 \kappa^4 \log d) \leq \frac{\mu k \kappa^2}{d}$$

Since initialization gives $\max_i g_i(\mathbf{U}_0) \leq \frac{10 \mu k \kappa^2}{d}$, therefore:

$$P(g_i(\mathbf{U}_t) 1_{\mathfrak{E}_{t-1}} \geq 20\frac{\mu k \kappa^2}{d}) \leq \frac{1}{2d^{11}}$$

That is equivalent to:

$$P(\mathfrak{E}_{t-1} \cap \{g_i(\mathbf{U}_t) \geq 20\frac{\mu k \kappa^2}{d}\}) \leq \frac{1}{2d^{11}} \tag{12}$$

Construction of supermartingale F: On the other hand, we also have

$$\begin{aligned}&\mathbb{E}\|SG(\mathbf{U}_t)\|_F^2 1_{\mathfrak{E}_t} \leq \mathbb{E}16d^4(\mathbf{u}_i^\top \mathbf{u}_j - \mathbf{M}_{ij})^2 \max_i \left\|\mathbf{e}_i^\top \mathbf{U}\right\|^2 1_{\mathfrak{E}_t}\\ \leq& 16d^2\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{M}\|_F^2 \max_i \left\|\mathbf{e}_i^\top \mathbf{U}_t\right\|^2 1_{\mathfrak{E}_t} \leq O(\mu dk\kappa^2) f(\mathbf{U}_t) 1_{\mathfrak{E}_t}\end{aligned}$$

Therefore, by update function Eq.(8),

$$\begin{aligned}&\mathbb{E}[f(\mathbf{U}_{t+1})1_{\mathfrak{E}_t}|\mathfrak{F}_t]\\ \leq&[f(\mathbf{U}_t) - \mathbb{E}\langle\nabla f(\mathbf{U}_t), \eta SG(\mathbf{U}_t)\rangle + \eta^2\mathbb{E}\left\|SG(\mathbf{U}_t)\right\|_F^2]1_{\mathfrak{E}_t}\\ =&[f(\mathbf{U}_t) - \eta\left\|\nabla f(\mathbf{U}_t)\right\|_F^2 + \eta^2\mathbb{E}\left\|SG(\mathbf{U}_t)\right\|_F^2]1_{\mathfrak{E}_t}\\ \leq&[(1-\frac{2\eta}{\kappa})f(\mathbf{U}_t) + \eta^2 O(\mu dk\kappa^2)f(\mathbf{U}_t)]1_{\mathfrak{E}_t}\\ \leq&(1-\frac{\eta}{\kappa})f(\mathbf{U}_t)1_{\mathfrak{E}_t}\end{aligned}$$

Let $F_t = (1-\frac{\eta}{\kappa})^{-t} f(\mathbf{U}_t)1_{\mathfrak{E}_{t-1}}$, we know F_t is also a supermartingale.

Probability 1 bound for F: With probabilty 1, we also have:

$$\begin{aligned}F_t - \mathbb{E}[F_t|\mathfrak{F}_{t-1}] =&(1-\frac{\eta}{\kappa})^{-t}[-\eta\langle\nabla f(\mathbf{U}_t), SG(\mathbf{U}_t) - \mathbb{E}SG(\mathbf{U}_t)\rangle\\ &+\frac{\eta^2}{2}(\nabla^2 f(\zeta_t)(SG(\mathbf{U}_t), SG(\mathbf{U}_t)) - \mathbb{E}\nabla^2 f(\zeta_t)(SG(\mathbf{U}_t), SG(\mathbf{U}_t)))]1_{\mathfrak{E}_{t-1}}\end{aligned} \tag{13}$$

where ζ_t depends on $SG(\mathbf{U}_t)$.

First, recall we denote $\text{SVD}(\mathbf{M}) = \mathbf{XSX}^\top$, and $\text{SVD}(\mathbf{UU}^\top) = \mathbf{WDW}^\top$, and observe that:

$$\begin{aligned}&\left\|\mathbf{UU}^\top - \mathbf{M}\right\|_\infty 1_{\mathfrak{E}_{t-1}} = \max_{ij}|\text{tr}(\mathbf{e}_i^\top(\mathbf{UU}^\top - \mathbf{M})\mathbf{e}_j)|1_{\mathfrak{E}_{t-1}}\\ =&\max_{ij}|\text{tr}(\mathbf{e}_i^\top(\mathcal{P}_{\mathbf{X}} + \mathcal{P}_{\mathbf{X}_\perp})(\mathbf{UU}^\top - \mathbf{M})\mathbf{e}_j)|1_{\mathfrak{E}_{t-1}}\\ \leq&\max_{ij}|\text{tr}(\mathbf{e}_i^\top\mathcal{P}_{\mathbf{X}}(\mathbf{UU}^\top - \mathbf{M})\mathbf{e}_j)|1_{\mathfrak{E}_{t-1}} + \max_{ij}|\text{tr}(\mathbf{e}_i^\top\mathcal{P}_{\mathbf{X}_\perp}\mathbf{UU}^\top\mathbf{e}_j)|1_{\mathfrak{E}_{t-1}}\\ \leq&\max_i\left\|\mathbf{e}_i^\top\mathbf{X}\right\|\left\|\mathbf{UU}^\top - \mathbf{M}\right\|_F 1_{\mathfrak{E}_{t-1}} + \max_i\left\|\mathbf{e}_i^\top\mathbf{W}\right\|\left\|\mathbf{UU}^\top - \mathbf{M}\right\|_F 1_{\mathfrak{E}_{t-1}}\\ \leq&O(\sqrt{\frac{\mu k\kappa^3}{d}})\sqrt{f(\mathbf{U}_t)}\end{aligned}$$

Then, when sample (i,j) entry of matrix $\mathbf{M}$, we have:

$$\begin{aligned}&\langle\nabla f(\mathbf{U}_t), SG(\mathbf{U}_t)\rangle 1_{\mathfrak{E}_{t-1}} \leq O(1)\left\|\nabla f(\mathbf{U}_t)\right\|_F\left\|SG(\mathbf{U}_t)\right\|_F 1_{\mathfrak{E}_{t-1}}\\ \leq&O(1)d^2\sqrt{f(\mathbf{U}_t)}(\mathbf{UU}^\top - \mathbf{M})_{ij}\left\|\mathbf{e}_i\mathbf{u}_j^\top + \mathbf{e}_j\mathbf{u}_i^\top\right\|_F^2 1_{\mathfrak{E}_{t-1}}\\ \leq&O(1)d^2\sqrt{f(\mathbf{U}_t)}\left\|\mathbf{UU}^\top - \mathbf{M}\right\|_\infty\max_i\left\|\mathbf{e}_i^\top\mathbf{U}\right\| 1_{\mathfrak{E}_{t-1}} \leq O(\mu dk\kappa^{2.5})f(\mathbf{U}_t)1_{\mathfrak{E}_{t-1}}\end{aligned}$$

and

$$\begin{aligned}&\nabla^2 f(\zeta_t)(SG(\mathbf{U}_t), SG(\mathbf{U}_t))1_{\mathfrak{E}_{t-1}} \leq O(1)\left\|SG(\mathbf{U}_t)\right\|_F^2\\ \leq&O(1)d^4\left\|\mathbf{UU}^\top - \mathbf{M}\right\|_\infty^2\max_i\left\|\mathbf{e}_i^\top\mathbf{U}\right\|^2 1_{\mathfrak{E}_{t-1}} \leq O(\mu^2d^2k^2\kappa^5)f(\mathbf{U}_t)1_{\mathfrak{E}_{t-1}}\end{aligned}$$

Therefore, by decomposition Eq.(13), we have with probability 1:

$$|F_t - \mathbb{E}[F_t|\mathfrak{F}_{t-1}]| \leq (1-\frac{\eta}{\kappa})^{-t}\eta O(\mu dk\kappa^{2.5})f(\mathbf{U}_{t-1})1_{\mathfrak{E}_{t-1}} \leq (1-\frac{\eta}{\kappa})^{-t}(1-\frac{\eta}{2\kappa})^t\eta O(\mu dk\kappa^{0.5})1_{\mathfrak{E}_{t-1}} \tag{14}$$

Variance bound for F: We also know

$$\begin{aligned}&\mathrm{Var}(\langle\nabla f(\mathbf{U}_t), SG(\mathbf{U}_t)\rangle 1_{\mathfrak{E}_{t-1}}|\mathfrak{F}_{t-1}) \le \mathbb{E}[\langle\nabla f(\mathbf{U}_t), SG(\mathbf{U}_t)\rangle^2 1_{\mathfrak{E}_{t-1}}|\mathfrak{F}_{t-1}]\\ \le& \|\nabla f(\mathbf{U}_t)\|_F^2 \,\mathbb{E}\,\|SG(\mathbf{U}_t)\|_F^2\, 1_{\mathfrak{E}_{t-1}} \le O(1)d^2 \max_i \left\|\mathbf{e}_i^\top \mathbf{U}\right\|^2 f^2(\mathbf{U}_{t-1}) 1_{\mathfrak{E}_{t-1}}\\ \le& O(\mu dk\kappa^2) f^2(\mathbf{U}_{t-1}) 1_{\mathfrak{E}_{t-1}}\end{aligned}$$

and

$$\begin{aligned}&\mathrm{Var}(\nabla^2 f(\zeta_t)(SG(\mathbf{U}_t), SG(\mathbf{U}_t)) 1_{\mathfrak{E}_{t-1}}|\mathfrak{F}_{t-1}) \le \mathbb{E}[\nabla^2 f(\zeta_t)(SG(\mathbf{U}_t), SG(\mathbf{U}_t))^2 1_{\mathfrak{E}_{t-1}}|\mathfrak{F}_{t-1}]\\ \le& O(1)\mathbb{E}\,\|SG(\mathbf{U}_t)\|_F^4 = O(1)\mathbb{E}d^8 (\mathbf{U}\mathbf{U}^\top - \mathbf{M})_{ij}^4 \max_i \left\|\mathbf{e}_i^\top \mathbf{U}\right\|^4 1_{\mathfrak{E}_{t-1}}\\ \le& O(1)d^6 \left\|\mathbf{U}\mathbf{U}^\top - \mathbf{M}\right\|_\infty^2 \left\|\mathbf{U}\mathbf{U}^\top - \mathbf{M}\right\|_F^2 \max_i \left\|\mathbf{e}_i^\top \mathbf{U}\right\|^4 1_{\mathfrak{E}_{t-1}}\\ \le& O(\mu^3 d^3 k^3 \kappa^7) f^2(\mathbf{U}_{t-1}) 1_{\mathfrak{E}_{t-1}}\end{aligned}$$

Therefore, by decomposition Eq.(13), we have:

$$\mathrm{Var}(F_t|\mathfrak{F}_{t-1}) \le (1-\frac{\eta}{\kappa})^{-2t}\eta^2 O(\mu dk\kappa^2) f^2(\mathbf{U}_{t-1}) 1_{\mathfrak{E}_{t-1}} \le (1-\frac{\eta}{\kappa})^{-2t}(1-\frac{\eta}{2\kappa})^{2t}\eta^2 O(\frac{\mu dk}{\kappa^2}) 1_{\mathfrak{E}_{t-1}} \tag{15}$$

Bernstein's inequality for F: Let $\sigma^2 = \sum_{\tau=1}^t \mathrm{Var}(F_\tau|\mathfrak{F}_{\tau-1})$, and R satisfies, with probability 1 that $|F_\tau - \mathbb{E}[F_\tau||\mathfrak{F}_{\tau-1}]| \le R,\ \tau = 1, \cdots, t$. Then By standard Bernstein concentration inequality, we know:

$$P(F_t \ge F_0 + s) \le \exp(\frac{s^2/2}{\sigma^2 + Rs/3})$$

Let $s' = O(1)(1-\frac{\eta}{\kappa})^t[\sqrt{\sigma^2 \log d} + R\log d]$, this gives:

$$P(f(\mathbf{U}_t)1_{\mathfrak{E}_{t-1}} \ge (1-\frac{\eta}{\kappa})^t f(\mathbf{U}_0) + s') \le \frac{1}{2d^{10}}$$

By Eq.(14), we know $R = (1-\frac{\eta}{\kappa})^{-t}(1-\frac{\eta}{2\kappa})^t \eta O(\mu dk\kappa^{0.5})$ satisfies that $|F_\tau - \mathbb{E}[F_\tau|\mathfrak{F}_{\tau-1}]| \le R,\ \tau = 1, \cdots, t$. Also by Eq. (15), we have:

$$\begin{aligned}&(1-\frac{\eta}{\kappa})^t\sqrt{\sigma^2\log d} \le \eta O(\sqrt{\frac{\mu dk\log d}{\kappa^2}})\sqrt{\sum_{\tau=1}^t (1-\frac{\eta}{\kappa})^{2t-2\tau}(1-\frac{\eta}{2\kappa})^{2\tau}}\\ \le& (1-\frac{\eta}{2\kappa})^t\eta O(\sqrt{\frac{\mu dk\log d}{\kappa^2}})\sqrt{\sum_{\tau=1}^t (1-\frac{\eta}{\kappa})^{2t-2\tau}(1-\frac{\eta}{2\kappa})^{2\tau-2t}} \le (1-\frac{\eta}{2\kappa})^t\sqrt{\eta}O(\sqrt{\frac{\mu dk\log d}{\kappa}})\end{aligned}$$

by $\eta < \frac{c}{\mu dk\kappa^3\log d}$ and choosing c to be small enough, we have:

$$s' = (1-\frac{\eta}{2\kappa})^t[\sqrt{\eta}O(\sqrt{\frac{\mu dk\log d}{\kappa}}) + \eta O(\mu dk\kappa^{0.5})] \le (1-\frac{\eta}{2\kappa})^t(\frac{1}{20\kappa})^2$$

Since $F_0 = f(\mathbf{U}_0) \le (\frac{1}{20\kappa})^2$, therefore:

$$P(f(\mathbf{U}_t)1_{\mathfrak{E}_{t-1}} \ge (1-\frac{\eta}{2\kappa})^t(\frac{1}{10\kappa})^2) \le \frac{1}{2d^{10}}$$

That is equivalent to:

$$P(\mathfrak{E}_{t-1} \cap \{f(\mathbf{U}_t) \ge (1-\frac{\eta}{2\kappa})^t(\frac{1}{10\kappa})^2\}) \le \frac{1}{2d^{10}} \tag{16}$$

Probability for event $\mathfrak{E}_T$: Finally, combining the concentration result for martingale G (Eq.(12)) and martingale F (Eq.(16)), we conclude:

$$\begin{aligned}&P(\mathfrak{E}_{t-1} \cap \bar{\mathfrak{E}}_t) = P\left[\mathfrak{E}_{t-1} \cap \left(\cup_i\{g_i(\mathbf{U}_t) \ge 20\frac{\mu k\kappa^2}{d}\} \cup \{f(\mathbf{U}_t) \ge (1-\frac{\eta}{2\kappa})^t(\frac{1}{10\kappa})^2\}\right)\right]\\ \le& \sum_{i=1}^d P(\mathfrak{E}_{t-1} \cap \{g_i(\mathbf{U}_t) \ge 20\frac{\mu k\kappa^2}{d}\}) + P(\mathfrak{E}_{t-1} \cap \{f(\mathbf{U}_t) \ge (1-\frac{\eta}{2\kappa})^t(\frac{1}{10\kappa})^2\}) \le \frac{1}{d^{10}}\end{aligned}$$

Since

$$P(\bar{\mathfrak{E}_T}) = \sum_{t=1}^{T} P(\mathfrak{E}_{t-1} \cap \bar{\mathfrak{E}_t}) \leq \frac{T}{d^{10}}$$

We finishes the proof. □

C Proof of General Asymmetric Case

In this section, we first prove Lemma 3.2, set up the equivalence between Algorithm 2 and Algorithm 3. Then we prove the main theorem for general asymmetric matrix (Theorem 3.3). WLOG, we continue to assume $\|\mathbf{M}\| = 1$ in all proof. Also, when it's clear from the context, we use κ to specifically to represent $\kappa(\mathbf{M})$. Then $\sigma_{\min}(\mathbf{M}) = \frac{1}{\kappa}$. Also in this section, we always use $d = \max\{d_1, d_2\}$ and denote $\text{SVD}(\mathbf{M}) = \mathbf{X}\mathbf{S}\mathbf{Y}^\top$, and $\text{SVD}(\mathbf{U}\mathbf{V}^\top) = \mathbf{W}_U\mathbf{D}\mathbf{W}_V^\top$.

Proof of Lemma 3.2. Let us always denote the iterates in Algorithm 2 by $\mathbf{U}_t$, $\mathbf{V}_t$, and denote the corresponding iterates in Algorithm 3 by $\mathbf{U}'_t$, $\mathbf{V}'_t$ using prime version. We use induction to prove the equivalence. Assume at time t we have $\mathbf{U}_t\mathbf{V}_t^\top = \mathbf{U}'_t\mathbf{V}_t'^\top$. Recall in Algorithm 2, we renormalize $\mathbf{U}_t, \mathbf{V}_t$ to $\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t$, this set up the correspondence:

$$\tilde{\mathbf{U}}_t = \mathbf{U}'_t\mathbf{R}'_U{\mathbf{D}'_U}^{-\frac{1}{2}}\mathbf{Q}'_U{\mathbf{D}'}^{\frac{1}{2}}$$
$$\tilde{\mathbf{V}}_t = \mathbf{V}'_t\mathbf{R}'_V{\mathbf{D}'_V}^{-\frac{1}{2}}\mathbf{Q}'_V{\mathbf{D}'}^{\frac{1}{2}}$$

Denote $\mathbf{P}'_U = \mathbf{R}'_U{\mathbf{D}'_U}^{-\frac{1}{2}}\mathbf{Q}'_U{\mathbf{D}'}^{\frac{1}{2}}$, and $\mathbf{P}'_V = \mathbf{V}'_t\mathbf{R}'_V{\mathbf{D}'_V}^{-\frac{1}{2}}\mathbf{Q}'_V{\mathbf{D}'}^{\frac{1}{2}}$. Clearly $\mathbf{P}'_U{\mathbf{P}'_V}^\top = \mathbf{I}$. Then we have $\tilde{\mathbf{U}}_t = \mathbf{U}'_t\mathbf{P}'_U$, $\tilde{\mathbf{V}}_t = \mathbf{P}'_V$ and thus:

$$\begin{aligned}
&\mathbf{U}_{t+1}\mathbf{V}_{t+1}^\top \\
=&(\tilde{\mathbf{U}}_t - 2\eta d_1 d_2(\tilde{\mathbf{U}}_t\tilde{\mathbf{V}}_t^\top - \mathbf{M})_{ij}\mathbf{e}_i\mathbf{e}_j^\top\tilde{\mathbf{V}}_t)(\tilde{\mathbf{V}}_t - 2\eta d_1 d_2(\tilde{\mathbf{U}}_t\tilde{\mathbf{V}}_t^\top - \mathbf{M})_{ij}\mathbf{e}_j\mathbf{e}_i^\top\tilde{\mathbf{U}}_t)^\top \\
=&(\mathbf{U}'_t\mathbf{P}'_U - 2\eta d_1 d_2(\mathbf{U}'_t\mathbf{V}_t'^\top - \mathbf{M})_{ij}\mathbf{e}_i\mathbf{e}_j^\top\mathbf{V}'_t\mathbf{P}'_V)(\mathbf{V}'_t\mathbf{P}'_V - 2\eta d_1 d_2(\mathbf{U}'_t\mathbf{V}_t'^\top - \mathbf{M})_{ij}\mathbf{e}_j\mathbf{e}_i^\top\mathbf{U}'_t\mathbf{P}'_U)^\top \\
=&(\mathbf{U}'_t - 2\eta d_1 d_2(\mathbf{U}'_t\mathbf{V}_t'^\top - \mathbf{M})_{ij}\mathbf{e}_i\mathbf{e}_j^\top\mathbf{V}'_t\mathbf{P}'_V{\mathbf{P}'_U}^{-1}) \\
&\cdot(\mathbf{V}'_t - 2\eta d_1 d_2(\mathbf{U}'_t\mathbf{V}_t'^\top - \mathbf{M})_{ij}\mathbf{e}_j\mathbf{e}_i^\top\mathbf{U}'_t\mathbf{P}'_U{\mathbf{P}'_V}^{-1})^\top \\
=&\mathbf{U}'_{t+1}{\mathbf{V}'_{t+1}}^\top
\end{aligned}$$

Clearly with same initialization algorithm, we have $\mathbf{U}_0\mathbf{V}_0^\top = \mathbf{U}'_0\mathbf{V}_0'^\top$, by induction, we finish the proof. □

Now we proceed to prove Theorem 3.3. Since Algorithm 2 and Algorithm 3 are equivalent, we will focus our analysis on Algorithm 2 which is more theoretical appealing. As for the symmetric PSD case, we first present the essential ingradient:

Theorem C.1. *Let* $f(\mathbf{U},\mathbf{V}) = \left\|\mathbf{U}\mathbf{V}^\top - \mathbf{M}\right\|_F^2$, $g_i(\mathbf{U},\mathbf{V}) = \left\|\mathbf{e}_i^\top\mathbf{U}\mathbf{V}^\top\right\|^2$, *and* $h_j(\mathbf{U},\mathbf{V}) = \left\|\mathbf{e}_j^\top\mathbf{V}\mathbf{U}^\top\right\|^2$, *for* $i \in [d_1]$ *and* $j \in [d_2]$. *Suppose after initialization, we have:*

$$f(\mathbf{U}_0,\mathbf{V}_0) \leq (\frac{1}{20\kappa})^2, \qquad \max_i g_i(\mathbf{U}_0,\mathbf{V}_0) \leq \frac{10\mu k\kappa^2}{d_1}, \qquad \max_j h_j(\mathbf{U}_0,\mathbf{V}_0) \leq \frac{10\mu k\kappa^2}{d_2}$$

Then, there exist some absolute constant c *such that for any learning rate* $\eta < \frac{c}{\mu dk\kappa^3\log d}$, *with at least* $1 - \frac{T}{d^{10}}$ *probability, we will have for all* $t \leq T$ *that:*

$$f(\mathbf{U}_t,\mathbf{V}_t) \leq (1 - \frac{\eta}{2\kappa})^t(\frac{1}{10\kappa})^2, \quad \max_i g_i(\mathbf{U}_t,\mathbf{V}_t) \leq \frac{100\mu k\kappa^2}{d_1}, \quad \max_j h_j(\mathbf{U}_t,\mathbf{V}_t) \leq \frac{100\mu k\kappa^2}{d_2}$$

Theorem 3.3 can easily be concluded from Theorem C.1 and Lemma A.1. Theorem C.1 also provides similar guarantees as Theorem B.1 in symmetric case. However, due to the additional invariance between $\mathbf{U}$ and $\mathbf{V}$, Theorem C.1 need to keep track of more complicated potential function $g_i(\mathbf{U},\mathbf{V})$ and $h_j(\mathbf{U},\mathbf{V})$ to control the incoherence, which makes the proof more involved.

The rest of this section all focus on proving Theorem C.1. Similar to symmetric PSD case, we also first prepare with a few lemmas about the property of objective function, and the spectral property of $\mathbf{U}, \mathbf{V}$ in a local Frobenius ball around optimal. Then, we prove Theorem C.1 by constructing three supermartingales related to $f(\mathbf{U}_t, \mathbf{V}_t), g_i(\mathbf{U}_t, \mathbf{V}_t), h_j(\mathbf{U}_t, \mathbf{V}_t)$ each, and applying concentration argument.

To make the notation clear, denote gradient $\nabla f(\mathbf{U}, \mathbf{V}) \in \mathbb{R}^{(d_1+d_2)\times k}$:

$$\nabla f(\mathbf{U}, \mathbf{V}) = \begin{pmatrix} \frac{\partial}{\partial \mathbf{U}} f(\mathbf{U}, \mathbf{V}) \\ \frac{\partial}{\partial \mathbf{V}} f(\mathbf{U}, \mathbf{V}) \end{pmatrix}$$

Also denote the stochastic gradient $SG(\mathbf{U}, \mathbf{V})$ by (if we sampled entry (i, j) of matrix $\mathbf{M}$)

$$SG(\mathbf{U}, \mathbf{V}) = 2d_1 d_2 (\mathbf{U}\mathbf{V}^\top - \mathbf{M})_{ij} \begin{pmatrix} \mathbf{e}_i \mathbf{e}_j^\top \mathbf{V} \\ \mathbf{e}_j \mathbf{e}_i^\top \mathbf{U} \end{pmatrix}$$

$$\mathbb{E}SG(\mathbf{U}, \mathbf{V}) = \nabla f(\mathbf{U}, \mathbf{V}) = 2 \begin{pmatrix} (\mathbf{U}\mathbf{V}^\top - \mathbf{M})\mathbf{V} \\ (\mathbf{U}\mathbf{V}^\top - \mathbf{M})^\top \mathbf{U} \end{pmatrix}$$

By update function, we know:

$$\begin{pmatrix} \mathbf{U}_{t+1} \\ \mathbf{V}_{t+1} \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{U}}_t \\ \tilde{\mathbf{V}}_t \end{pmatrix} - \eta SG(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t)$$

and $\tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top = \mathbf{U}_t \mathbf{V}_t^\top$ is the renormalized version of $\mathbf{U}_t \mathbf{V}_t^\top$.

C.1 Geometric Properties in Local Region

Similar to symmetric PSD case, we first prove two lemmas w.r.t the smoothness and property similar to strongly convex for objective function:

Lemma C.2. *Within the region $\mathcal{D} = \{(\mathbf{U}, \mathbf{V}) | \, \|\mathbf{U}\| \le \Gamma, \|\mathbf{V}\| \le \Gamma\}$, we have function $f(\mathbf{U}, \mathbf{V}) = \|\mathbf{M} - \mathbf{U}\mathbf{V}^\top\|_F^2$ satisfying:*

$$\|\nabla f(\mathbf{U}_1, \mathbf{V}_1) - \nabla f(\mathbf{U}_2, \mathbf{V}_2)\|_F^2 \le \beta^2 (\|\mathbf{U}_1 - \mathbf{U}_2\|_F^2 + \|\mathbf{V}_1 - \mathbf{V}_2\|_F^2)$$

where smoothness parameter $\beta = 8\max\{\Gamma^2, \|\mathbf{M}\|\}$.

Proof. Inside region $\mathcal{D}$, we have:

$$\begin{aligned}
&\|\nabla f(\mathbf{U}_1, \mathbf{V}_1) - \nabla f(\mathbf{U}_2, \mathbf{V}_2)\|_{\mathrm{F}}^2 \\
=&\|\frac{\partial}{\partial \mathbf{U}} f(\mathbf{U}_1, \mathbf{V}_1) - \frac{\partial}{\partial \mathbf{U}} f(\mathbf{U}_2, \mathbf{V}_2)\|_{\mathrm{F}}^2 + \|\frac{\partial}{\partial \mathbf{V}} f(\mathbf{U}_1, \mathbf{V}_1) - \frac{\partial}{\partial \mathbf{V}} f(\mathbf{U}_2, \mathbf{V}_2)\|_{\mathrm{F}}^2 \\
=&4(\|(\mathbf{U}_1\mathbf{V}_1^\top - \mathbf{M})\mathbf{V}_1 - (\mathbf{U}_2\mathbf{V}_2^\top - \mathbf{M})\mathbf{V}_2\|_{\mathrm{F}}^2 + \|(\mathbf{U}_1\mathbf{V}_1^\top - \mathbf{M})^\top \mathbf{U}_1 - (\mathbf{U}_2\mathbf{V}_2^\top - \mathbf{M})^\top \mathbf{U}_2\|_{\mathrm{F}}^2) \\
\le&64\max\{\Gamma^4, \|\mathbf{M}\|^2\}(\|\mathbf{U}_1 - \mathbf{U}_2\|_{\mathrm{F}}^2 + \|\mathbf{V}_1 - \mathbf{V}_2\|_{\mathrm{F}}^2)
\end{aligned}$$

The last step is by similar technics as in the proof of Lemma B.2, by expanding

$$\mathbf{U}_1\mathbf{V}_1^\top\mathbf{V}_1 - \mathbf{U}_2\mathbf{V}_2^\top\mathbf{V}_2 = (\mathbf{U}_1 - \mathbf{U}_2)\mathbf{V}_1^\top\mathbf{V}_1 + \mathbf{U}_2(\mathbf{V}_1 - \mathbf{V}_2)^\top\mathbf{V}_1 + \mathbf{U}_2\mathbf{V}_2^\top(\mathbf{V}_1 - \mathbf{V}_2)$$

□

Lemma C.3. *Within the region $\mathcal{D} = \{(\mathbf{U}, \mathbf{V}) | \sigma_{\min}(\mathbf{X}^\top\mathbf{U}) \ge \gamma, \sigma_{\min}(\mathbf{Y}^\top\mathbf{V}) \ge \gamma\}$, then we have function $f(\mathbf{U}, \mathbf{V}) = \|\mathbf{M} - \mathbf{U}\mathbf{V}^\top\|_F^2$ satisfying:*

$$\|\nabla f(\mathbf{U}, \mathbf{V})\|_F^2 \ge \alpha f(\mathbf{U}, \mathbf{V})$$

where constant $\alpha = 4\gamma^2$.

Proof. Let $\hat{\mathbf{U}}, \hat{\mathbf{V}}$ be the left singular vectors of $\mathbf{U}, \mathbf{V}$. Inside region $\mathcal{D}$, we have:

$$\begin{aligned}
&\|(\mathbf{U}\mathbf{V}^\top - \mathbf{M})\mathbf{V}\|_{\mathrm{F}}^2 \\
=&\|\mathcal{P}_{\hat{\mathbf{U}}}(\mathbf{U}\mathbf{V}^\top - \mathbf{M})\mathbf{V}\|_{\mathrm{F}}^2 + \|\mathcal{P}_{\hat{\mathbf{U}}_\perp}(\mathbf{U}\mathbf{V}^\top - \mathbf{M})\mathbf{V}\|_{\mathrm{F}}^2 \\
\ge&\sigma_k(\mathbf{V})^2\|\mathcal{P}_{\hat{\mathbf{U}}}(\mathbf{U}\mathbf{V}^\top - \mathbf{M})\mathcal{P}_{\hat{\mathbf{V}}}\|_{\mathrm{F}}^2 + \|\mathcal{P}_{\hat{\mathbf{U}}_\perp}\mathbf{M}\mathbf{V}\|_{\mathrm{F}}^2 \\
\ge&\sigma_k(\mathbf{V})^2\|\mathcal{P}_{\hat{\mathbf{U}}}(\mathbf{U}\mathbf{V}^\top - \mathbf{M})\mathcal{P}_{\hat{\mathbf{V}}}\|_{\mathrm{F}}^2 + \sigma_{\min}(\mathbf{Y}^\top\mathbf{V})^2\|\mathcal{P}_{\hat{\mathbf{U}}_\perp}\mathbf{X}\mathbf{\Sigma}\|_{\mathrm{F}}^2 \\
=&\sigma_k(\mathbf{V})^2\|\mathcal{P}_{\hat{\mathbf{U}}}(\mathbf{U}\mathbf{V}^\top - \mathbf{M})\mathcal{P}_{\hat{\mathbf{V}}}\|_{\mathrm{F}}^2 + \sigma_{\min}(\mathbf{Y}^\top\mathbf{V})^2\|\mathcal{P}_{\hat{\mathbf{U}}_\perp}\mathbf{M}\|_{\mathrm{F}}^2
\end{aligned}$$

Therefore, by symmetry, we have:

$$\begin{aligned}\|\nabla f(\mathbf{U},\mathbf{V})\|_F^2 =& 4(\|(\mathbf{U}\mathbf{V}^\top-\mathbf{M})\mathbf{V}\|_{\mathrm{F}}^2+\|(\mathbf{U}\mathbf{V}^\top-\mathbf{M})^\top\mathbf{U}\|_{\mathrm{F}}^2)\\ \geq& 4\gamma^2(2\|\mathcal{P}_{\hat{\mathbf{U}}}(\mathbf{U}\mathbf{V}^\top-\mathbf{M})\mathcal{P}_{\hat{\mathbf{V}}}\|_{\mathrm{F}}^2+\|\mathcal{P}_{\hat{\mathbf{U}}_\perp}\mathbf{M}\|_{\mathrm{F}}^2+\|\mathbf{M}\mathcal{P}_{\hat{\mathbf{V}}_\perp}\|_{\mathrm{F}}^2)\\ \geq& 4\gamma^2\|\mathbf{U}\mathbf{V}^\top-\mathbf{M}\|_{\mathrm{F}}^2\end{aligned}$$

□

Next, we show as long as we are in some Frobenious ball around optimum, then we have good spectral property over $\mathbf{U},\mathbf{V}$ which guarantees the preconditions for Lemma C.2 and Lemma C.3.

Lemma C.4. *Within the region* $\mathcal{D}=\{(\mathbf{U},\mathbf{V})|\left\|\mathbf{M}-\mathbf{U}\mathbf{V}^\top\right\|_F\leq\frac{1}{10}\sigma_k(\mathbf{M})\}$*, and for* $\mathbf{U}=\mathbf{W}_U\mathbf{D}^{\frac{1}{2}},\mathbf{V}=\mathbf{W}_V\mathbf{D}^{\frac{1}{2}}$ *where* $\mathbf{W}_U\mathbf{D}\mathbf{W}_V=SVD(\mathbf{U}\mathbf{V}^\top)$*, we have:*

$$\begin{aligned}\|\mathbf{U}\|\leq\sqrt{2\|\mathbf{M}\|},\qquad \sigma_{\min}(\mathbf{X}^\top\mathbf{U})\geq\sqrt{\sigma_k(\mathbf{M})/2}\\ \|\mathbf{V}\|\leq\sqrt{2\|\mathbf{M}\|},\qquad \sigma_{\min}(\mathbf{Y}^\top\mathbf{V})\geq\sqrt{\sigma_k(\mathbf{M})/2}\end{aligned}$$

Proof. For spectral norm of $\mathbf{U}$, we have:

$$\|\mathbf{U}\|^2=\|\mathbf{D}\|=\left\|\mathbf{U}\mathbf{V}^\top\right\|\leq\|\mathbf{M}\|+\left\|\mathbf{M}-\mathbf{U}\mathbf{V}^\top\right\|\leq\|\mathbf{M}\|+\left\|\mathbf{M}-\mathbf{U}\mathbf{V}^\top\right\|_F\leq 2\|\mathbf{M}\|$$

For the minimum singular value of $\mathbf{U}^\top\mathbf{U}$, we have:

$$\begin{aligned}\sigma_{\min}(\mathbf{U}^\top\mathbf{U})=&\sigma_k(\mathbf{D})=\sigma_k(\mathbf{U}\mathbf{V}^\top)\geq\sigma_k(\mathbf{M})-\left\|\mathbf{M}-\mathbf{U}\mathbf{V}^\top\right\|\\ \geq&\sigma_k(\mathbf{M})-\left\|\mathbf{M}-\mathbf{U}\mathbf{U}^\top\right\|_F\geq\frac{9}{10}\sigma_k(\mathbf{M})\end{aligned}$$

By symmetry, the same holds for $\mathbf{V}$. On the other hand, we have:

$$\begin{aligned}\frac{1}{10}\sigma_k(\mathbf{M})\geq&\left\|\mathbf{M}-\mathbf{U}\mathbf{V}^\top\right\|_F\geq\left\|\mathcal{P}_{\mathbf{X}_\perp}(\mathbf{M}-\mathbf{U}\mathbf{V}^\top)\right\|_F=\left\|\mathcal{P}_{\mathbf{X}_\perp}\mathbf{U}\mathbf{V}^\top\right\|_F=\|\mathcal{P}_{\mathbf{X}_\perp}\mathbf{W}_U\mathbf{D}\|_F\\ \geq&\|\mathcal{P}_{\mathbf{X}_\perp}\mathbf{W}_U\mathbf{D}\|\geq\frac{9}{10}\sigma_k(\mathbf{M})\|\mathbf{X}_\perp\mathbf{W}_U\|\end{aligned}$$

Let the principal angle between $\mathbf{X}$ and $\mathbf{W}_U$ to be θ. This gives $\sin^2\theta=\left\|\mathbf{X}_\perp^\top\mathbf{W}_U\right\|^2\leq\frac{1}{9}$. Thus $\cos^2\theta=\sigma_{\min}^2(\mathbf{X}^\top\mathbf{W}_U)\geq\frac{8}{9}$. Therefore:

$$\sigma_{\min}^2(\mathbf{X}^\top\mathbf{U})\geq\sigma_{\min}^2(\mathbf{X}^\top\mathbf{W}_U)\sigma_{\min}(\mathbf{U}^\top\mathbf{U})\geq\sigma_k(\mathbf{M})/2$$

□

C.2 Proof of Theorem C.1

Now, we are ready for our key theorem. By Lemma C.2, Lemma C.3, and Lemma C.4, we already know the function has good property locally in the region $\mathcal{D}=\{(\mathbf{U},\mathbf{V})|\left\|\mathbf{M}-\mathbf{U}\mathbf{V}^\top\right\|_F\leq\frac{1}{10}\sigma_k(\mathbf{M})\}$ which alludes linear convergence. Similar to the symmetric PSD case, the work remains is to prove that once we initialize inside this region, our algorithm will guarantee $\mathbf{U},\mathbf{V}$ never leave this region with high probability even with relatively large stepsize. Again, we also need to control the incoherence of $\mathbf{U}_t,\mathbf{V}_t$ over all iterates additionally to achieve tight sample complexity and near optimal runtime.

Following is our formal proof.

Proof of Theorem C.1. For simplicity of notation, we assume $d=d_1=d_2$, and do not distinguish d_1 and d_2. However, it is easy to check our proof never use the property $\mathbf{M}$ is square matrix. The proof easily extends to $d_1\neq d_2$ case by replacing d in the proof with suitable d_1,d_2.

Define event $\mathfrak{E}_t=\{\forall\tau\leq t,f(\mathbf{U}_\tau,\mathbf{V}_\tau)\leq(1-\frac{\eta}{2\kappa})^t(\frac{1}{10\kappa})^2,\max_i g_i(\mathbf{U}_\tau,\mathbf{V}_\tau)\leq\frac{100\mu k\kappa^2}{d},\max_j h_j(\mathbf{U}_\tau,\mathbf{V}_\tau)\leq\frac{100\mu k\kappa^2}{d}\}$. Theorem C.1 is equivalent to prove event $\mathfrak{E}_T$ happens with high probability. The proof achieves this by contructing two supermartingales for $f(\mathbf{U}_t,\mathbf{V}_t)1_{\mathfrak{E}_t}$, $g_i(\mathbf{U}_t,\mathbf{V}_t)1_{\mathfrak{E}_t}$ and $h_i(\mathbf{U}_t,\mathbf{V}_t)1_{\mathfrak{E}_t}$ (where $1_{(\cdot)}$ denote indicator function), applies concentration argument.

The proofs also follow similar structure as symmetric PSD case:

1. The constructions of supermartingales
2. Their probability 1 bound and variance bound in order to apply Azuma-Bernstein inequality
3. Final combination of concentration results to conclude the proof

Then let filtration $\mathfrak{F}_t = \sigma\{SG(\mathbf{U}_0, \mathbf{V}_0), \cdots, SG(\mathbf{U}_{t-1}, \mathbf{V}_{t-1})\}$ where $\sigma\{\cdot\}$ denotes the sigma field. Also let event , note $\mathfrak{E}_t \subset \mathfrak{F}_t$. Also $\mathfrak{E}_{t+1} \subset \mathfrak{E}_t$, and thus $1_{\mathfrak{E}_{t+1}} \leq 1_{\mathfrak{E}_t}$.

By Lemma C.4, we immediately know that conditioned on $\mathfrak{E}_t$, we have $\|\mathbf{U}_t\| \leq \sqrt{2}$, $\|\mathbf{V}_t\| \leq \sqrt{2}$, $\sigma_{\min}(\mathbf{X}^\top \mathbf{U}_t) \geq 1/\sqrt{2\kappa}$, $\sigma_{\min}(\mathbf{Y}^\top \mathbf{V}_t) \geq 1/\sqrt{2\kappa}$. We will use this fact throughout the proof.

For simplicity, when it's clear from the context, we denote:

$$\begin{pmatrix} \Delta_{\mathbf{U}} \\ \Delta_{\mathbf{V}} \end{pmatrix} = -\eta SG(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t) = \begin{pmatrix} \mathbf{U}_{t+1} \\ \mathbf{V}_{t+1} \end{pmatrix} - \begin{pmatrix} \tilde{\mathbf{U}}_t \\ \tilde{\mathbf{V}}_t \end{pmatrix}$$

Construction of supermartingale G: First, since potential function $g_l(\mathbf{U}, \mathbf{V})$ is forth-order polynomial, we can expand:

$$\begin{aligned} g_l(\tilde{\mathbf{U}}_{t+1}, \tilde{\mathbf{V}}_{t+1}) &= g_l(\mathbf{U}_{t+1}, \mathbf{V}_{t+1}) = g_l(\tilde{\mathbf{U}}_t + \Delta_{\mathbf{U}}, \tilde{\mathbf{V}}_t + \Delta_{\mathbf{V}}) \\ &= \mathbf{e}_l^\top (\tilde{\mathbf{U}}_t + \Delta_{\mathbf{U}})(\tilde{\mathbf{V}}_t + \Delta_{\mathbf{V}})^\top (\tilde{\mathbf{V}}_t + \Delta_{\mathbf{V}})(\tilde{\mathbf{U}}_t + \Delta_{\mathbf{U}})^\top \mathbf{e}_l \\ &= g_l(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t) + 2\mathbf{e}_l^\top \Delta_{\mathbf{U}} \tilde{\mathbf{V}}_t^\top \tilde{\mathbf{V}}_t \tilde{\mathbf{U}}_t^\top \mathbf{e}_l + 2\mathbf{e}_l^\top \tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top \Delta_{\mathbf{V}} \tilde{\mathbf{U}}_t \mathbf{e}_l + R_2 \\ &= g_l(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t) + R_1 \end{aligned}$$

Where we denote R_1 as the sum of first order terms and higher order terms (all second/third/forth order terms), and R_2 as the sum of second order terms and higher order terms.

We now give a proposition about properties of R_1 and R_2 which involves a lot calculation, and postpone its proof in the end of this section.

Proposition C.5. *With above notations, we have following inequalities hold true.*

$$\begin{aligned} &\mathbb{E}[R_2 1_{\mathfrak{E}_t} | \mathfrak{F}_t] \leq \eta^2 O(\mu^2 k^2 \kappa^4) 1_{\mathfrak{E}_t} \\ &|R_1| 1_{\mathfrak{E}_t} \leq \eta O(\mu^2 k^2 \kappa^5) 1_{\mathfrak{E}_t} \qquad \textit{w.p } 1 \\ &\mathbb{E}[R_1^2 1_{\mathfrak{E}_t} | \mathfrak{F}_t] \leq \eta^2 O(\frac{\mu^3 k^3 \kappa^6}{d}) 1_{\mathfrak{E}_t} \end{aligned}$$

Then by taking conditional expectation, we have:

$$\begin{aligned} &\mathbb{E}[g_l(\tilde{\mathbf{U}}_{t+1}, \tilde{\mathbf{V}}_{t+1}) 1_{\mathfrak{E}_t} | \mathfrak{F}_t] = \mathbb{E}[g_l(\mathbf{U}_{t+1}, \mathbf{V}_{t+1}) 1_{\mathfrak{E}_t} | \mathfrak{F}_t] \\ \leq &\mathbb{E}[g_l(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t) + 2\mathbf{e}_l^\top \Delta_{\mathbf{U}} \tilde{\mathbf{V}}_t^\top \tilde{\mathbf{V}}_t \tilde{\mathbf{U}}_t^\top \mathbf{e}_l + 2\mathbf{e}_l^\top \tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top \Delta_{\mathbf{V}} \tilde{\mathbf{U}}_t \mathbf{e}_l + R_2] 1_{\mathfrak{E}_t} \end{aligned}$$

The first order term can be calculated as:

$$\begin{aligned} &[-\mathbb{E} 2\mathbf{e}_l^\top \Delta_{\mathbf{U}} \tilde{\mathbf{V}}_t^\top \tilde{\mathbf{V}}_t \tilde{\mathbf{U}}_t^\top \mathbf{e}_l + 2\mathbf{e}_l^\top \tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top \Delta_{\mathbf{V}} \tilde{\mathbf{U}}_t \mathbf{e}_l] 1_{\mathfrak{E}_t} \\ =&[-4\mathbf{e}_l^\top (\tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M}) \tilde{\mathbf{V}}_t \tilde{\mathbf{V}}_t^\top \tilde{\mathbf{V}}_t \tilde{\mathbf{U}}_t^\top \mathbf{e}_l - 4\mathbf{e}_l^\top \tilde{\mathbf{U}}_t \tilde{\mathbf{U}}_t^\top (\tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M}^\top) \tilde{\mathbf{V}}_t \tilde{\mathbf{U}}_t^\top \mathbf{e}_l] 1_{\mathfrak{E}_t} \\ =&[-4\mathbf{e}_l^\top \tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top \tilde{\mathbf{V}}_t \tilde{\mathbf{V}}_t^\top \tilde{\mathbf{V}}_t \tilde{\mathbf{U}}_t^\top \mathbf{e}_l + 4\mathbf{e}_l^\top \mathbf{M} \tilde{\mathbf{V}}_t \tilde{\mathbf{V}}_t^\top \tilde{\mathbf{V}}_t \tilde{\mathbf{U}}_t^\top \mathbf{e}_l - 4\mathbf{e}_l^\top \tilde{\mathbf{U}}_t \tilde{\mathbf{U}}_t^\top (\tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M}^\top) \tilde{\mathbf{V}}_t \tilde{\mathbf{U}}_t^\top \mathbf{e}_l] 1_{\mathfrak{E}_t} \\ \leq &[-4[\sigma_{\min}(\tilde{\mathbf{V}}_t^\top \tilde{\mathbf{V}}_t) \left\| \mathbf{e}_l^\top \tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top \right\|^2 + \left\| \mathbf{e}_l^\top \tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top \right\| \left\| \tilde{\mathbf{V}}_t \tilde{\mathbf{V}}_t^\top \right\| \left\| \mathbf{e}_l^\top \mathbf{M} \right\| \\ &+ \left\| \mathbf{e}_l^\top \tilde{\mathbf{U}}_t \tilde{\mathbf{U}}_t^\top \right\| \left\| \tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M}^\top \right\|_F \left\| \mathbf{e}_l^\top \tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top \right\|]] 1_{\mathfrak{E}_t} \\ \leq &[-\frac{2}{\kappa} g_l(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t) + 80\frac{\mu k \kappa}{d} + \frac{4}{10\kappa} g_l(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t)] 1_{\mathfrak{E}_t} \\ \leq &[-\frac{1}{\kappa} g_l(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t) + 80\frac{\mu k \kappa}{d}] 1_{\mathfrak{E}_t} \end{aligned}$$

In second last inequality, we use key observation:

$$\left\| \mathbf{e}_k^\top \tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top \right\| = \left\| \mathbf{e}_k^\top \mathbf{W}_{\mathbf{U}} \mathbf{D} \mathbf{W}_{\mathbf{V}}^\top \right\| = \left\| \mathbf{e}_k^\top \mathbf{W}_{\mathbf{U}} \mathbf{D} \mathbf{W}_{\mathbf{U}}^\top \right\| = \left\| \mathbf{e}_k^\top \tilde{\mathbf{U}}_t \tilde{\mathbf{U}}_t^\top \right\|$$

By Proposition C.5, we know $\mathbb{E}[R_2 1_{\mathfrak{E}_t}|\mathfrak{F}_t] \leq \eta^2 O(\mu^2 k^2 \kappa^4) 1_{\mathfrak{E}_t}$. Combine both facts and recall $\eta < \frac{c}{\mu d k \kappa^3 \log d}$, we have:

$$\begin{aligned}\mathbb{E}[g_i(\tilde{\mathbf{U}}_{t+1}, \tilde{\mathbf{V}}_{t+1}) 1_{\mathfrak{E}_t}|\mathfrak{F}_t] &\leq [(1-\frac{\eta}{\kappa}) g_i(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t) + \frac{80\eta\mu k\kappa}{d} + O(\eta^2\mu^2 k^2 \kappa^4)] 1_{\mathfrak{E}_t} \\ &\leq [(1-\frac{\eta}{\kappa}) g_i(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t) + \frac{90\eta\mu k\kappa}{d}] 1_{\mathfrak{E}_t}\end{aligned}$$

The last inequality is achieved by choosing c small enough.

Let $G_{it} = (1-\frac{\eta}{\kappa})^{-t}(g_i(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t) 1_{\mathfrak{E}_{t-1}} - 90\frac{\mu k\kappa^2}{d})$. This gives:

$$\mathbb{E}[G_{i(t+1)}|\mathfrak{F}_t] \leq (1-\frac{\eta}{\kappa})^{-t}(g_i(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t) 1_{\mathfrak{E}_t} - 90\frac{\mu k\kappa^2}{d}) \leq G_{it}$$

The right inequality is true since $1_{\mathfrak{E}_t} \leq 1_{\mathfrak{E}_{t-1}}$. This implies G_{it} is supermartingale.

Probability 1 bound for G: We also know:

$$G_{i(t+1)} - \mathbb{E}[G_{i(t+1)}|\mathfrak{F}_t] = (1-\frac{\eta}{\kappa})^{-(t+1)}[R_1 - \mathbb{E}R_1] 1_{\mathfrak{E}_t}$$

By Proposition C.5, we know with probability 1 that $|R_1| 1_{\mathfrak{E}_t} \leq \eta O(\mu^2 k^2 \kappa^5) 1_{\mathfrak{E}_t}$. This gives with probability 1:

$$|G_{it} - \mathbb{E}[G_{it}|\mathfrak{F}_{t-1}]| \leq (1-\frac{\eta}{\kappa})^{-t} \eta O(\mu^2 k^2 \kappa^5) 1_{\mathfrak{E}_{t-1}} \tag{17}$$

Variance bound for G: We also know

$$\mathrm{Var}(G_{i(t+1)}|\mathfrak{F}_t) = (1-\frac{\eta}{\kappa})^{-2(t+1)}[\mathbb{E}R_1^2 1_{\mathfrak{E}_t} - (\mathbb{E}R_1 1_{\mathfrak{E}_t})^2] \leq \mathbb{E}[R_1^2 1_{\mathfrak{E}_t}|\mathfrak{F}_t]$$

By Proposition C.5, we know that $\mathbb{E}[R_1^2 1_{\mathfrak{E}_t}|\mathfrak{F}_t] \leq \eta^2 O(\frac{\mu^3 k^3 \kappa^6}{d}) 1_{\mathfrak{E}_t}$. This gives:

$$\mathrm{Var}(G_{it}|\mathfrak{F}_{t-1}) \leq (1-\frac{\eta}{\kappa})^{-2t} \eta^2 O(\frac{\mu^3 k^3 \kappa^6}{d}) 1_{\mathfrak{E}_{t-1}} \tag{18}$$

Bernstein's inequality for G: Let $\sigma^2 = \sum_{\tau=1}^t \mathrm{Var}(G_{i\tau}|\mathfrak{F}_{\tau-1})$, and R satisfies, with probability 1 that $|G_{i\tau} - \mathbb{E}[G_{i\tau}|\mathfrak{F}_{\tau-1}]| \leq R,\ \tau = 1, \cdots, t$. Then By standard Bernstein concentration inequality, we know:

$$P(G_{it} \geq G_{i0} + s) \leq \exp(\frac{s^2/2}{\sigma^2 + Rs/3})$$

Since $G_{i0} = g_i(\tilde{\mathbf{U}}_0, \tilde{\mathbf{V}}_0) - 90\frac{\mu k\kappa^2}{d}$, let $s' = O(1)(1-\frac{\eta}{\kappa})^t[\sqrt{\sigma^2 \log d} + R\log d]$, we know

$$P\left(g_i(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t) 1_{\mathfrak{E}_{t-1}} \geq 90\frac{\mu k\kappa^2}{d} + (1-\frac{\eta}{\kappa})^t (g_i(\tilde{\mathbf{U}}_0, \tilde{\mathbf{V}}_0) - 90\frac{\mu k\kappa^2}{d}) + s'\right) \leq \frac{1}{3d^{11}}$$

By Eq.(17), we know $R = (1-\frac{\eta}{\kappa})^{-t}\eta O(\mu^2 k^2 \kappa^5)$ satisfies that $|G_{i\tau} - \mathbb{E}[G_{i\tau}|\mathfrak{F}_{\tau-1}]| \leq R,\ \tau = 1, \cdots, t$. Also by Eq. (18), we have:

$$(1-\frac{\eta}{\kappa})^t \sqrt{\sigma^2 \log d} \leq \eta O(\sqrt{\frac{\mu^3 k^3 \kappa^6 \log d}{d}}) \sqrt{\sum_{\tau=1}^t (1-\frac{\eta}{\kappa})^{2t-2\tau}} \leq \sqrt{\eta} O(\sqrt{\frac{\mu^3 k^3 \kappa^7 \log d}{d}})$$

by $\eta < \frac{c}{\mu d k \kappa^3 \log d}$ and choosing c to be small enough, we have:

$$s' = \sqrt{\eta} O(\sqrt{\frac{\mu^3 k^3 \kappa^7 \log d}{d}}) + \eta O(\mu^2 k^2 \kappa^5 \log d) \leq 10\frac{\mu k\kappa^2}{d}$$

Since initialization gives $\max_i g_i(\mathbf{U}_0, \mathbf{V}_0) \leq \frac{10\mu k\kappa^2}{d}$, therefore:

$$P(g_i(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t) 1_{\mathfrak{E}_{t-1}} \geq 100\frac{\mu k\kappa^2}{d}) \leq \frac{1}{3d^{11}}$$

That is equivalent to:

$$P(\mathfrak{E}_{t-1} \cap \{g_i(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t) \geq 100\frac{\mu k\kappa^2}{d}\}) \leq \frac{1}{3d^{11}} \tag{19}$$

By symmetry, we can also have corresponding result for $h_j(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t)$.

Construction of supermartingale F: Similarly, we also need to construct a martingale for $f(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t)$. Again, we can write f as forth order polynomial:

$$\begin{aligned} f(\tilde{\mathbf{U}}_{t+1}, \tilde{\mathbf{V}}_{t+1}) &= f(\mathbf{U}_{t+1}, \mathbf{V}_{t+1}) = f(\tilde{\mathbf{U}}_t + \Delta_{\mathbf{U}}, \tilde{\mathbf{V}}_t + \Delta_{\mathbf{V}}) \\ &= \text{tr}\left([(\tilde{\mathbf{U}}_t + \Delta_{\mathbf{U}})(\tilde{\mathbf{V}}_t + \Delta_{\mathbf{V}}) - \mathbf{M}][(\tilde{\mathbf{U}}_t + \Delta_{\mathbf{U}})(\tilde{\mathbf{V}}_t + \Delta_{\mathbf{V}}) - \mathbf{M}]^\top\right) \\ &= f(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t) + 2\text{tr}(\Delta_{\mathbf{U}}\tilde{\mathbf{V}}_t^\top(\tilde{\mathbf{U}}_t\tilde{\mathbf{V}}_t^\top - \mathbf{M})^\top) + 2\text{tr}(\Delta_{\mathbf{V}}\tilde{\mathbf{U}}_t^\top(\tilde{\mathbf{U}}_t\tilde{\mathbf{V}}_t^\top - \mathbf{M})) + Q_2 \\ &= f(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t) + Q_1 \end{aligned}$$

Where we denote Q_1 as the sum of first order terms and higher order terms (all second/third/forth order terms), and Q_2 as the sum of second order terms and higher order terms.

We also now give a proposition about properties of Q_1 and Q_2 which involves a lot calculation, and postpone its proof in the end of this section.

Proposition C.6. *With above notations, we have following inequalities hold true.*

$$\begin{aligned} &\mathbb{E}[Q_2 1_{\mathfrak{E}_t}|\mathfrak{F}_t] \leq \eta^2 O(\mu dk\kappa^2) f(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t) 1_{\mathfrak{E}_t} \\ &|Q_1|1_{\mathfrak{E}_t} \leq \eta O(\mu dk\kappa^3) f(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t) 1_{\mathfrak{E}_t} \qquad w.p\ 1 \\ &\mathbb{E}[Q_1^2 1_{\mathfrak{E}_t}|\mathfrak{F}_t] \leq \eta^2 O(\mu dk\kappa^2) f^2(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t) 1_{\mathfrak{E}_t} \end{aligned}$$

Then by Proposition C.6, we know $\mathbb{E}[Q_2 1_{\mathfrak{E}_t}|\mathfrak{F}_t] \leq \eta^2 O(\mu dk\kappa^2) f(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t) 1_{\mathfrak{E}_t}$. By taking conditional expectation, we have:

$$\begin{aligned} &\mathbb{E}[f(\mathbf{U}_{t+1})1_{\mathfrak{E}_t}|\mathfrak{F}_t] \\ \leq &[f(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t) - \mathbb{E}\langle \nabla f(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t), \eta SG(\mathbf{U}_t)\rangle + \mathbb{E}Q_2]1_{\mathfrak{E}_t} \\ = &[f(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t) - \eta \left\|\nabla f(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t)\right\|_F^2 + \mathbb{E}Q_2]1_{\mathfrak{E}_t} \\ \leq &[(1 - \frac{2\eta}{\kappa}) f(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t) + \eta^2 O(\mu dk\kappa^2) f(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t)]1_{\mathfrak{E}_t} \\ \leq &(1 - \frac{\eta}{\kappa}) f(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t) 1_{\mathfrak{E}_t} \end{aligned}$$

Let $F_t = (1 - \frac{\eta}{\kappa})^{-t} f(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t) 1_{\mathfrak{E}_{t-1}}$, we know F_t is also a supermartingale.

Probability 1 bound: We also know

$$F_{t+1} - \mathbb{E}[F_{t+1}|\mathfrak{F}_t] = (1 - \frac{\eta}{\kappa})^{-(t+1)}[Q_1 - \mathbb{E}Q_1]1_{\mathfrak{E}_t}$$

By Proposition C.6, we know with probability 1 that $|Q_1|1_{\mathfrak{E}_t} \leq \eta O(\mu dk\kappa^3) f(\mathbf{U}_t, \mathbf{V}_t) 1_{\mathfrak{E}_t}$. This gives with probability 1:

$$|F_t - \mathbb{E}F_t| \leq (1 - \frac{\eta}{\kappa})^{-t} \eta O(\mu dk\kappa^3) f(\mathbf{U}_{t-1}) 1_{\mathfrak{E}_{t-1}} \leq (1 - \frac{\eta}{\kappa})^{-t} (1 - \frac{\eta}{2\kappa})^t \eta O(\mu dk\kappa) 1_{\mathfrak{E}_{t-1}} \quad (20)$$

Variance bound: We also know

$$\text{Var}(F_{t+1}|\mathfrak{F}_t) = (1 - \frac{\eta}{\kappa})^{-2(t+1)}[\mathbb{E}Q_1^2 1_{\mathfrak{E}_t} - (\mathbb{E}Q_1 1_{\mathfrak{E}_t})^2] \leq (1 - \frac{\eta}{\kappa})^{-2(t+1)} \mathbb{E}[Q_1^2 1_{\mathfrak{E}_t}|\mathfrak{F}_t]$$

By Proposition C.6, we know that $\mathbb{E}[Q_1^2 1_{\mathfrak{E}_t}|\mathfrak{F}_t] \leq \eta^2 O(\mu dk\kappa^2) f^2(\mathbf{U}_t, \mathbf{V}_t) 1_{\mathfrak{E}_t}$. This gives:

$$\text{Var}(F_t|\mathfrak{F}_{t-1}) \leq (1 - \frac{\eta}{\kappa})^{-2t} \eta^2 O(\mu dk\kappa^2) f^2(\mathbf{U}_{t-1}) 1_{\mathfrak{E}_{t-1}} \leq (1 - \frac{\eta}{\kappa})^{-2t} (1 - \frac{\eta}{2\kappa})^{2t} \eta^2 O(\frac{\mu dk}{\kappa^2}) 1_{\mathfrak{E}_{t-1}} \quad (21)$$

Bernstein's inequality: Let $\sigma^2 = \sum_{\tau=1}^{t} \mathrm{Var}(F_\tau|\mathfrak{F}_{\tau-1})$, and R satisfies, with probability 1 that $|F_\tau - \mathbb{E}[F_\tau||\mathfrak{F}_{\tau-1}]| \le R,\ \tau = 1, \cdots, t$. Then By standard Bernstein concentration inequality, we know:

$$P(F_t \ge F_0 + s) \le \exp(\frac{s^2/2}{\sigma^2 + Rs/3})$$

Let $s' = O(1)(1-\frac{\eta}{\kappa})^t[\sqrt{\sigma^2 \log d} + R\log d]$, this gives:

$$P(f(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t)1_{\mathfrak{E}_{t-1}} \ge (1-\frac{\eta}{\kappa})^t f(\mathbf{U}_0) + s') \le \frac{1}{3d^{10}}$$

By Eq.(20), we know $R = (1-\frac{\eta}{\kappa})^{-t}(1-\frac{\eta}{2\kappa})^t \eta O(\mu dk\kappa)$ satisfies that $|F_\tau - \mathbb{E}[F_\tau|\mathfrak{F}_{\tau-1}]| \le R,\ \tau = 1, \cdots, t$. Also by Eq. (21), we have:

$$(1-\frac{\eta}{\kappa})^t\sqrt{\sigma^2\log d} \le \eta O(\sqrt{\frac{\mu dk\log d}{\kappa^2}})\sqrt{\sum_{\tau=1}^{t}(1-\frac{\eta}{\kappa})^{2t-2\tau}(1-\frac{\eta}{2\kappa})^{2\tau}}$$

$$\le (1-\frac{\eta}{2\kappa})^t \eta O(\sqrt{\frac{\mu dk\log d}{\kappa^2}})\sqrt{\sum_{\tau=1}^{t}(1-\frac{\eta}{\kappa})^{2t-2\tau}(1-\frac{\eta}{2\kappa})^{2\tau-2t}} \le (1-\frac{\eta}{2\kappa})^t\sqrt{\eta}O(\sqrt{\frac{\mu dk\log d}{\kappa}})$$

by $\eta < \frac{c}{\mu dk\kappa^3\log d}$ and choosing c to be small enough, we have:

$$s' = (1-\frac{\eta}{2\kappa})^t[\sqrt{\eta}O(\sqrt{\frac{\mu dk\kappa\log d}{\kappa}}) + \eta O(\mu dk\kappa)] \le (1-\frac{\eta}{2\kappa})^t(\frac{1}{20\kappa})^2$$

Since $F_0 = f(\mathbf{U}_0) \le (\frac{1}{20\kappa})^2$, therefore:

$$P(f(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t)1_{\mathfrak{E}_{t-1}} \ge (1-\frac{\eta}{2\kappa})^t(\frac{1}{10\kappa})^2) \le \frac{1}{3d^{10}}$$

That is equivalent to:

$$P(\mathfrak{E}_{t-1} \cap \{f(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t) \ge (1-\frac{\eta}{2\kappa})^t(\frac{1}{10\kappa})^2\}) \le \frac{1}{3d^{10}} \tag{22}$$

Probability for event $\mathfrak{E}_T$: Finally, combining the concentration result for martingale G (Eq.(19)) and martingale F (Eq.(22)), we conclude:

$$\begin{aligned}
&P(\mathfrak{E}_{t-1} \cap \bar{\mathfrak{E}}_t) \\
=&P\left[\mathfrak{E}_{t-1} \cap \left([\cup_i\{g_i(\mathbf{U}_t, \mathbf{V}_t) \ge 100\frac{\mu k\kappa^2}{d}\}]\right.\right. \\
&\left.\left. \cup [\cup_j\{h_j(\mathbf{U}_t, \mathbf{V}_t) \ge 100\frac{\mu k\kappa^2}{d}\}] \cup \{f(\mathbf{U}_t) \ge (1-\frac{\eta}{2\kappa})^t(\frac{1}{10\kappa})^2\}\right)\right] \\
\le& 2\sum_{i=1}^{d} P(\mathfrak{E}_{t-1} \cap \{g_i(\mathbf{U}_t, \mathbf{V}_t) \ge 100\frac{\mu k\kappa^2}{d}\}) + P(\mathfrak{E}_{t-1} \cap \{f(\mathbf{U}_t, \mathbf{V}_t) \ge (1-\frac{\eta}{2\kappa})^t(\frac{1}{10\kappa})^2\}) \\
\le& \frac{1}{d^{10}}
\end{aligned}$$

Since

$$P(\bar{\mathfrak{E}_T}) = \sum_{t=1}^{T} P(\mathfrak{E}_{t-1} \cap \bar{\mathfrak{E}}_t) \le \frac{T}{d^{10}}$$

We finishes the proof. □

Finally we give proof for Proposition C.5 and Proposition C.6. The proof mostly consistsof expanding every term and careful calculations.

Proof of Proposition C.5. For simplicity of notation, we hide the term $1_{\mathfrak{E}_t}$ in all following equations. Reader should always think every term in this proof multiplied by $1_{\mathfrak{E}_t}$. Recall that:

$$SG(\mathbf{U},\mathbf{V}) = 2d^2(\mathbf{U}\mathbf{V}^\top - \mathbf{M})_{ij}\begin{pmatrix}\mathbf{e}_i\mathbf{e}_j^\top\mathbf{V}\\ \mathbf{e}_j\mathbf{e}_i^\top\mathbf{U}\end{pmatrix}$$
$$\begin{pmatrix}\Delta_{\mathbf{U}}\\ \Delta_{\mathbf{V}}\end{pmatrix} = -\eta SG(\tilde{\mathbf{U}}_t,\tilde{\mathbf{V}}_t) = \begin{pmatrix}\mathbf{U}_{t+1}\\ \mathbf{V}_{t+1}\end{pmatrix} - \begin{pmatrix}\tilde{\mathbf{U}}_t\\ \tilde{\mathbf{V}}_t\end{pmatrix}$$

We first prove first three inequality. Recall that:

$$\begin{aligned}g_l(\tilde{\mathbf{U}}_{t+1},\tilde{\mathbf{V}}_{t+1}) &= g_l(\mathbf{U}_{t+1},\mathbf{V}_{t+1}) = g_l(\tilde{\mathbf{U}}_t+\Delta_{\mathbf{U}},\tilde{\mathbf{V}}_t+\Delta_{\mathbf{V}})\\ &= \mathbf{e}_l^\top(\tilde{\mathbf{U}}_t+\Delta_{\mathbf{U}})(\tilde{\mathbf{V}}_t+\Delta_{\mathbf{V}})^\top(\tilde{\mathbf{V}}_t+\Delta_{\mathbf{V}})(\tilde{\mathbf{U}}_t+\Delta_{\mathbf{U}})^\top\mathbf{e}_l\\ &= g_l(\tilde{\mathbf{U}}_t,\tilde{\mathbf{V}}_t) + 2\mathbf{e}_l^\top\Delta_{\mathbf{U}}\tilde{\mathbf{V}}_t^\top\tilde{\mathbf{V}}_t\tilde{\mathbf{U}}_t^\top\mathbf{e}_l + 2\mathbf{e}_l^\top\tilde{\mathbf{U}}_t\tilde{\mathbf{V}}_t^\top\Delta_{\mathbf{V}}\tilde{\mathbf{U}}_t\mathbf{e}_l + R_2\\ &= g_l(\tilde{\mathbf{U}}_t,\tilde{\mathbf{V}}_t) + R_1\end{aligned}$$

By expanding the polynomial, we can write out the first order term:

$$\begin{aligned}R_1 - R_2 =& 2\mathbf{e}_l^\top\Delta_{\mathbf{U}}\tilde{\mathbf{V}}_t^\top\tilde{\mathbf{V}}_t\tilde{\mathbf{U}}_t^\top\mathbf{e}_l + 2\mathbf{e}_l^\top\tilde{\mathbf{U}}_t\tilde{\mathbf{V}}_t^\top\Delta_{\mathbf{V}}\tilde{\mathbf{U}}_t\mathbf{e}_l\\ =& -4\eta d^2(\tilde{\mathbf{U}}_t\tilde{\mathbf{V}}_t^\top - \mathbf{M})_{ij}\left(\delta_{il}\mathbf{e}_j^\top\tilde{\mathbf{V}}_t\tilde{\mathbf{V}}_t^\top\tilde{\mathbf{V}}_t\tilde{\mathbf{U}}_t^\top\mathbf{e}_l + (\tilde{\mathbf{U}}_t\tilde{\mathbf{V}}_t^\top)_{lj}(\tilde{\mathbf{U}}_t\tilde{\mathbf{U}}_t^\top)_{il}\right)\end{aligned}$$

Second order term:

$$\begin{aligned}&R_2 - R_3\\ =&\mathbf{e}_l^\top\Delta_{\mathbf{U}}\tilde{\mathbf{V}}_t^\top\tilde{\mathbf{V}}_t\Delta_{\mathbf{U}}^\top\mathbf{e}_l + \mathbf{e}_l^\top\tilde{\mathbf{U}}_t\Delta_{\mathbf{V}}^\top\Delta_{\mathbf{V}}\tilde{\mathbf{U}}_t^\top\mathbf{e}_l + 2\mathbf{e}_l^\top\Delta_{\mathbf{U}}\tilde{\mathbf{V}}_t^\top\Delta_{\mathbf{V}}\tilde{\mathbf{U}}_t^\top\mathbf{e}_l + 2\mathbf{e}_l^\top\Delta_{\mathbf{U}}\Delta_{\mathbf{V}}^\top\tilde{\mathbf{V}}_t\tilde{\mathbf{U}}_t^\top\mathbf{e}_l\\ =&4\eta^2d^4(\tilde{\mathbf{U}}_t\tilde{\mathbf{V}}_t^\top - \mathbf{M})_{ij}^2\\ &\cdot\left(\delta_{il}\left\|\mathbf{e}_j^\top\tilde{\mathbf{V}}_t\tilde{\mathbf{V}}_t^\top\right\|^2 + (\tilde{\mathbf{U}}_t\tilde{\mathbf{U}}_t^\top)_{li}^2 + 2\delta_{il}(\tilde{\mathbf{V}}_t\tilde{\mathbf{V}}_t^\top)_{jj}(\tilde{\mathbf{U}}_t\tilde{\mathbf{U}}_t^\top)_{ii} + 2\delta_{il}(\tilde{\mathbf{U}}_t\tilde{\mathbf{V}}_t^\top)_{ij}^2\right)\end{aligned}$$

Third order term:

$$\begin{aligned}R_4 - R_3 =& 2\mathbf{e}_l^\top\Delta_{\mathbf{U}}\tilde{\mathbf{V}}_t^\top\Delta_{\mathbf{V}}\Delta_{\mathbf{U}}^\top\mathbf{e}_l + 2\mathbf{e}_l^\top\Delta_{\mathbf{U}}\Delta_{\mathbf{V}}^\top\Delta_{\mathbf{V}}\tilde{\mathbf{U}}_t^\top\mathbf{e}_l\\ =& -16\eta^3d^6(\tilde{\mathbf{U}}_t\tilde{\mathbf{V}}_t^\top - \mathbf{M})_{ij}^3\delta_{il}\left((\tilde{\mathbf{V}}_t\tilde{\mathbf{V}}_t)_{jj}(\tilde{\mathbf{U}}_t\tilde{\mathbf{V}}_t)_{ij} + (\tilde{\mathbf{U}}_t\tilde{\mathbf{V}}_t)_{ij}(\tilde{\mathbf{U}}_t\tilde{\mathbf{U}}_t)_{ii}\right)\end{aligned}$$

Fourth order term:

$$R_4 = \mathbf{e}_l^\top\Delta_{\mathbf{U}}\Delta_{\mathbf{V}}^\top\Delta_{\mathbf{V}}\Delta_{\mathbf{U}}^\top\mathbf{e}_l = 16\eta^4d^8(\tilde{\mathbf{U}}_t\tilde{\mathbf{V}}_t^\top - \mathbf{M})_{ij}^4\delta_{il}(\tilde{\mathbf{U}}_t\tilde{\mathbf{V}}_t)_{ij}^2$$

For the ease of proof, we denote $\chi = \frac{\mu k\kappa^2}{d}$, then we know conditioned on event $\mathfrak{E}_t$, we have: $\max_i\left\|\mathbf{e}_i^\top\tilde{\mathbf{U}}_t\tilde{\mathbf{V}}_t^\top\right\|^2 \le O(\chi)$, and $\max_j\left\|\mathbf{e}_j^\top\tilde{\mathbf{V}}_t\tilde{\mathbf{U}}_t^\top\right\|^2 \le O(\chi)$. Some key inequality we need to use in the proof are listed here:

$$\left\|\mathbf{e}_l^\top\tilde{\mathbf{U}}_t\tilde{\mathbf{V}}_t^\top\right\| = \left\|\mathbf{e}_l^\top\tilde{\mathbf{U}}_t\tilde{\mathbf{U}}_t^\top\right\| \quad\text{and}\quad \left\|\mathbf{e}_l^\top\tilde{\mathbf{V}}_t\tilde{\mathbf{U}}_t^\top\right\| = \left\|\mathbf{e}_l^\top\tilde{\mathbf{V}}_t\tilde{\mathbf{V}}_t^\top\right\| \tag{23}$$

and

$$|(\tilde{\mathbf{U}}_t\tilde{\mathbf{V}}_t)_{ij}| \le \left\|\mathbf{e}_i^\top\tilde{\mathbf{U}}_t\tilde{\mathbf{V}}_t^\top\right\| \le O(\sqrt{\chi}) \tag{24}$$

The same also holds true for:

$$|(\tilde{\mathbf{U}}_t\tilde{\mathbf{U}}_t^\top)_{ii}| \le O(\sqrt{\chi}) \quad\text{and}\quad |(\tilde{\mathbf{V}}_t\tilde{\mathbf{V}}_t^\top)_{jj}| \le O(\sqrt{\chi}) \tag{25}$$

Another fact we frequently used is:

$$\frac{1}{2}\left\|\mathbf{e}_i^\top\mathbf{U}\mathbf{V}^\top\right\|^2 \le \left\|\mathbf{e}_i^\top\mathbf{U}\right\|^2 \le 2\kappa\left\|\mathbf{e}_i^\top\mathbf{U}\mathbf{V}^\top\right\|^2$$

This gives:

$$\left\|\tilde{\mathbf{U}}_t\tilde{\mathbf{V}}_t^\top - \mathbf{M}\right\|_\infty \le \max_k\left\|\mathbf{e}_k\tilde{\mathbf{U}}_t\right\|\max_k\left\|\mathbf{e}_k\tilde{\mathbf{V}}_t\right\| + \max_k\|\mathbf{e}_k\mathbf{X}\|\max_k\|\mathbf{e}_k\mathbf{Y}\|\,\|\mathbf{S}\| \le O(\chi\kappa)$$

and recall we choose $\eta < \frac{c}{\mu dk\kappa^3\log d}$, where c is some universal constant, then we have:

$$\eta d^2\left\|\tilde{\mathbf{U}}_t\tilde{\mathbf{V}}_t^\top - \mathbf{M}\right\|_\infty = O(\eta d^2\chi\kappa) \le O(1) \tag{26}$$

With equation (23), (24), (25), (26), now we are ready to prove Lemma.

For the first inequality $\mathbb{E}[R_2 1_{\mathfrak{E}_t}|\mathfrak{F}_t] \leq \eta^2 O(\mu^2 k^2 \kappa^4) 1_{\mathfrak{E}_t}$:

$$\mathbb{E}[R_2 1_{\mathfrak{E}_t}|\mathfrak{F}_t] \leq \mathbb{E}[|R_2 - R_3|1_{\mathfrak{E}_t}|\mathfrak{F}_t] + \mathbb{E}[|R_3 - R_4|1_{\mathfrak{E}_t}|\mathfrak{F}_t] + \mathbb{E}[|R_4|1_{\mathfrak{E}_t}|\mathfrak{F}_t]$$

For each term, we can bound as:

$$\begin{aligned}
\mathbb{E}[|R_2 - R_3|1_{\mathfrak{E}_t}|\mathfrak{F}_t] \leq & \eta^2 O(d^2) \sum_{ij} (\tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M})_{ij}^2 \left(\delta_{il} O(\chi) + (\tilde{\mathbf{U}}_t \tilde{\mathbf{U}}_t^\top)_{li}^2 \right) \\
\leq & \eta^2 O(d^2) \max_{l'} \left\| \mathbf{e}_{l'}^\top (\tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M}) \right\|^2 \sum_i \left(\delta_{il} O(\chi) + (\tilde{\mathbf{U}}_t \tilde{\mathbf{U}}_t^\top)_{li}^2 \right) \leq \eta^2 O(d^2 \chi^2) \\
\mathbb{E}[|R_3 - R_4|1_{\mathfrak{E}_t}|\mathfrak{F}_t] \leq & \eta^3 O(d^4) \sum_{ij} |\tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M}|_{ij}^3 \delta_{il} O(\chi) \\
\leq & \eta^3 O(d^4) \left\| \tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M} \right\|_\infty \left\| \mathbf{e}_l^\top (\tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M}) \right\|^2 O(\chi) \leq \eta^2 O(d^2 \chi^2) \\
\mathbb{E}[|R_4|1_{\mathfrak{E}_t}|\mathfrak{F}_t] \leq & \eta^4 O(d^6) \sum_{ij} (\tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M})_{ij}^4 \delta_{il} O(\chi) \\
\leq & \eta^4 O(d^6) \left\| \tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M} \right\|_\infty^2 \left\| \mathbf{e}_l^\top (\tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M}) \right\|^2 O(\chi) \leq \eta^2 O(d^2 \chi^2)
\end{aligned}$$

This gives in sum that

$$\mathbb{E}[R_2 1_{\mathfrak{E}_t}|\mathfrak{F}_t] \leq \eta^2 O(d^2 \chi^2) 1_{\mathfrak{E}_t} = \eta^2 O(\mu d k \kappa^2) f^2(\mathbf{U}_t) 1_{\mathfrak{E}_t}$$

For the second inequality $|R_1|1_{\mathfrak{E}_t} \leq \eta O(\mu^2 k^2 \kappa^5) 1_{\mathfrak{E}_t}$ w.p 1:

$$|R_1|1_{\mathfrak{E}_t} \leq |R_1 - R_2|1_{\mathfrak{E}_t} + |R_2 - R_3|1_{\mathfrak{E}_t} + |R_3 - R_4|1_{\mathfrak{E}_t} + |R_4|1_{\mathfrak{E}_t}$$

For each term, we can bound as:

$$\begin{aligned}
|R_1 - R_2|1_{\mathfrak{E}_t} \leq & \eta O(d^2) \left\| \tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M} \right\|_\infty O(\chi) \leq \eta O(d^2 \chi^2 \kappa) \\
|R_2 - R_3|1_{\mathfrak{E}_t} \leq & \eta^2 O(d^4) \left\| \tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M} \right\|_\infty^2 O(\chi) \leq \eta O(d^2 \chi^2 \kappa) \\
|R_3 - R_4|1_{\mathfrak{E}_t} \leq & \eta^3 O(d^6) \left\| \tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M} \right\|_\infty^3 O(\chi) \leq \eta O(d^2 \chi^2 \kappa) \\
|R_4|1_{\mathfrak{E}_t} \leq & \eta^4 O(d^8) \left\| \tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M} \right\|_\infty^4 O(\chi) \leq \eta O(d^2 \chi^2 \kappa)
\end{aligned}$$

This gives in sum that, with probability 1:

$$|R_1|1_{\mathfrak{E}_t} \leq \eta O(d^2 \chi^2 \kappa) 1_{\mathfrak{E}_t} = \eta O(\mu^2 k^2 \kappa^5) 1_{\mathfrak{E}_t}$$

For the third inequality $\mathbb{E}[R_1^2 1_{\mathfrak{E}_t}|\mathfrak{F}_t] \leq \eta^2 O(\frac{\mu^3 k^3 \kappa^6}{d}) 1_{\mathfrak{E}_t}$:

$$\mathbb{E} R_1^2 1_{\mathfrak{E}_t} \leq 4 \left[\mathbb{E}(R_1 - R_2)^2 1_{\mathfrak{E}_t} + \mathbb{E}(R_2 - R_3)^2 1_{\mathfrak{E}_t} + \mathbb{E}(R_3 - R_4)^2 1_{\mathfrak{E}_t} + \mathbb{E} R_4^2 1_{\mathfrak{E}_t} \right]$$

For each term, we can bound as:

$$\mathbb{E}(R_1 - R_2)^2 1_{\mathfrak{E}_t} \le \eta^2 O(d^2) \sum_{ij} (\tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M})_{ij}^2 \left(\delta_{il} O(\chi^2) + (\tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top)_{lj}^2 (\tilde{\mathbf{U}}_t \tilde{\mathbf{U}}_t^\top)_{il}^2 \right)$$

$$\le \eta^2 O(d^2) \max_{l'} \left\| \mathbf{e}_{l'}^\top (\tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M}) \right\|^2 \sum_i \left(\delta_{il} O(\chi^2) + (\tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top)_{lj}^2 (\tilde{\mathbf{U}}_t \tilde{\mathbf{U}}_t^\top)_{il}^2 \right) \le \eta^2 O(d^2 \chi^3)$$

$$\mathbb{E}(R_2 - R_3)^2 1_{\mathfrak{E}_t} \le \eta^4 O(d^6) \sum_{ij} (\tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M})_{ij}^4 \left(\delta_{il} O(\chi^2) + (\tilde{\mathbf{U}}_t \tilde{\mathbf{U}}_t^\top)_{li}^4 \right)$$

$$\le \eta^4 O(d^6) \left\| \tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M} \right\|_\infty^2 \max_{l'} \left\| \mathbf{e}_{l'}^\top (\tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M}) \right\|^2 \sum_i \left(\delta_{il} O(\chi^2) + (\tilde{\mathbf{U}}_t \tilde{\mathbf{U}}_t^\top)_{li}^4 \right) \le \eta^2 O(d^2 \chi^3)$$

$$\mathbb{E}(R_3 - R_4)^2 1_{\mathfrak{E}_t} \le \eta^6 O(d^{10}) \sum_{ij} |\tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M}|_{ij}^6 \delta_{il} O(\chi^2)$$

$$\le \eta^6 O(d^{10}) \left\| \tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M} \right\|_\infty^4 \left\| \mathbf{e}_l^\top (\tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M}) \right\|^2 O(\chi^2) \le \eta^2 O(d^2 \chi^3)$$

$$\mathbb{E} R_4^2 1_{\mathfrak{E}_t} \le \eta^8 O(d^{14}) \sum_{ij} (\tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M})_{ij}^8 \delta_{il} O(\chi^2)$$

$$\le \eta^8 O(d^{14}) \left\| \tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M} \right\|_\infty^6 \left\| \mathbf{e}_l^\top (\tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M}) \right\|^2 O(\chi^2) \le \eta^2 O(d^2 \chi^3)$$

This gives in sum that:

$$\mathbb{E} R_1^2 1_{\mathfrak{E}_t} \le \eta^2 O(d^2 \chi^3) 1_{\mathfrak{E}_t} = \eta^2 O(\frac{\mu^3 k^3 \kappa^6}{d}) 1_{\mathfrak{E}_t}$$

This finishes the proof. □

Proof of Proposition C.6. Similarly to the proof of Proposition C.5, we hide the term $1_{\mathfrak{E}_t}$ in all following equations. Reader should always think every term in this proof multiplied by $1_{\mathfrak{E}_t}$. Recall that:

$$\begin{aligned} f(\tilde{\mathbf{U}}_{t+1}, \tilde{\mathbf{V}}_{t+1}) &= f(\mathbf{U}_{t+1}, \mathbf{V}_{t+1}) = f(\tilde{\mathbf{U}}_t + \Delta_{\mathbf{U}}, \tilde{\mathbf{V}}_t + \Delta_{\mathbf{V}}) \\ &= \text{tr}\left([(\tilde{\mathbf{U}}_t + \Delta_{\mathbf{U}})(\tilde{\mathbf{V}}_t + \Delta_{\mathbf{V}}) - \mathbf{M}][(\tilde{\mathbf{U}}_t + \Delta_{\mathbf{U}})(\tilde{\mathbf{V}}_t + \Delta_{\mathbf{V}}) - \mathbf{M}]^\top \right) \\ &= f(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t) + 2\text{tr}(\Delta_{\mathbf{U}} \tilde{\mathbf{V}}_t^\top (\tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M})^\top) + 2\text{tr}(\Delta_{\mathbf{V}} \tilde{\mathbf{U}}_t^\top (\tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M})) + Q_2 \\ &= f(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t) + Q_1 \end{aligned}$$

By expanding the polynomial, we can write out the first order term:

$$\begin{aligned} Q_1 - Q_2 =& 2\text{tr}(\Delta_{\mathbf{U}} \tilde{\mathbf{V}}_t^\top (\tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M})^\top) + 2\text{tr}(\Delta_{\mathbf{V}} \tilde{\mathbf{U}}_t^\top (\tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M})) \\ =& -4\eta d^2 (\mathbf{U}\mathbf{V}^\top - \mathbf{M})_{ij} \left(\mathbf{e}_j^\top \tilde{\mathbf{V}}_t \tilde{\mathbf{V}}_t^\top (\tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M})^\top \mathbf{e}_i + \mathbf{e}_i^\top \tilde{\mathbf{U}}_t \tilde{\mathbf{U}}_t^\top (\tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M}) \mathbf{e}_j \right) \end{aligned}$$

The second order term:

$$\begin{aligned} & Q_2 - Q_3 \\ =& \text{tr}(\Delta_{\mathbf{U}} \tilde{\mathbf{V}}_t^\top \tilde{\mathbf{V}}_t \Delta_{\mathbf{U}}^\top) + \text{tr}(\tilde{\mathbf{U}}_t \Delta_{\mathbf{V}}^\top \Delta_{\mathbf{V}} \tilde{\mathbf{U}}_t^\top) + 2\text{tr}(\Delta_{\mathbf{U}} \tilde{\mathbf{V}}_t^\top \Delta_{\mathbf{V}} \tilde{\mathbf{U}}_t^\top) + 2\text{tr}(\Delta_{\mathbf{U}} \Delta_V^\top (\tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M})^\top) \\ =& 4\eta^2 d^4 (\mathbf{U}\mathbf{V}^\top - \mathbf{M})_{ij}^2 \\ & \cdot \left(\left\| \mathbf{e}_j^\top \tilde{\mathbf{V}}_t \tilde{\mathbf{V}}_t^\top \right\|^2 + \left\| \mathbf{e}_i^\top \tilde{\mathbf{U}}_t \tilde{\mathbf{U}}_t^\top \right\|^2 + (\tilde{\mathbf{V}}_t \tilde{\mathbf{V}}_t^\top)_{jj} (\tilde{\mathbf{U}}_t \tilde{\mathbf{U}}_t^\top)_{ii} + (\tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top)_{ij} (\tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M})_{ij} \right) \end{aligned}$$

The third order term:

$$\begin{aligned} Q_3 - Q_4 =& 2\text{tr}(\Delta_{\mathbf{U}} \tilde{\mathbf{V}}_t^\top \Delta_{\mathbf{V}} \Delta_{\mathbf{U}}^\top) + 2\text{tr}(\Delta_{\mathbf{U}} \Delta_{\mathbf{V}}^\top \Delta_{\mathbf{V}} \tilde{\mathbf{U}}_t^\top) \\ =& -16\eta^3 d^6 (\mathbf{U}\mathbf{V}^\top - \mathbf{M})_{ij}^3 \left((\tilde{\mathbf{V}}_t \tilde{\mathbf{V}}_t)_{jj} (\tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t)_{ij} + (\tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t)_{ij} (\tilde{\mathbf{U}}_t \tilde{\mathbf{U}}_t)_{ii} \right) \end{aligned}$$

The forth order term:

$$Q_4 = \text{tr}(\Delta_{\mathbf{U}}\Delta_{\mathbf{V}}^\top\Delta_{\mathbf{V}}\Delta_{\mathbf{U}}^\top) = 16\eta^4 d^8(\mathbf{U}\mathbf{V}^\top - \mathbf{M})_{ij}^4(\tilde{\mathbf{U}}_t\tilde{\mathbf{V}}_t)_{ij}^2$$

Again, in addition to equation (23), (23), (24), (25), we also need following inequality:

$$\begin{aligned}
&\left\|\tilde{\mathbf{U}}_t\tilde{\mathbf{V}}_t^\top - \mathbf{M}\right\|_\infty 1_{\mathfrak{E}_t} = \max_{ij}|\text{tr}(\mathbf{e}_i^\top(\tilde{\mathbf{U}}_t\tilde{\mathbf{V}}_t^\top - \mathbf{M})\mathbf{e}_j)|1_{\mathfrak{E}_t}\\
=&\max_{ij}|\text{tr}(\mathbf{e}_i^\top(\mathcal{P}_{\mathbf{X}} + \mathcal{P}_{\mathbf{X}_\perp})(\tilde{\mathbf{U}}_t\tilde{\mathbf{V}}_t^\top - \mathbf{M})\mathbf{e}_j)|1_{\mathfrak{E}_t}\\
\le&\max_{ij}|\text{tr}(\mathbf{e}_i^\top\mathcal{P}_{\mathbf{X}}(\tilde{\mathbf{U}}_t\tilde{\mathbf{V}}_t^\top - \mathbf{M})\mathbf{e}_j)|1_{\mathfrak{E}_t} + \max_{ij}|\text{tr}(\mathbf{e}_i^\top\mathcal{P}_{\mathbf{X}_\perp}\tilde{\mathbf{U}}_t\tilde{\mathbf{V}}_t^\top\mathbf{e}_j)|1_{\mathfrak{E}_t}\\
\le&\max_i\left\|\mathbf{e}_i^\top\mathbf{X}\right\|\left\|\tilde{\mathbf{U}}_t\tilde{\mathbf{V}}_t^\top - \mathbf{M}\right\|_F 1_{\mathfrak{E}_t} + \max_j\left\|\mathbf{e}_j^\top\mathbf{W}_{\mathbf{V}}\right\|\left\|\tilde{\mathbf{U}}_t\tilde{\mathbf{V}}_t^\top - \mathbf{M}\right\|_F 1_{\mathfrak{E}_t}\\
\le&O(\kappa\sqrt{\chi})\sqrt{f(\mathbf{U}_t)}
\end{aligned} \tag{27}$$

Now we are ready to prove Lemma.

For the first inequality $\mathbb{E}[Q_2 1_{\mathfrak{E}_t}|\mathfrak{F}_t] \le \eta^2 O(\mu dk\kappa^2)f(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t)1_{\mathfrak{E}_t}$**:**

$$\mathbb{E}[Q_2 1_{\mathfrak{E}_t}|\mathfrak{F}_t] \le \mathbb{E}[|Q_2 - Q_3|1_{\mathfrak{E}_t}|\mathfrak{F}_t] + \mathbb{E}[|Q_3 - Q_4|1_{\mathfrak{E}_t}|\mathfrak{F}_t] + \mathbb{E}[|Q_4|1_{\mathfrak{E}_t}|\mathfrak{F}_t]$$

For each term, we can bound as:

$$\begin{aligned}
\mathbb{E}[|Q_2 - Q_3|1_{\mathfrak{E}_t}|\mathfrak{F}_t] \le&\eta^2 O(d^2)\sum_{ij}(\tilde{\mathbf{U}}_t\tilde{\mathbf{V}}_t^\top - \mathbf{M})_{ij}^2 O(\chi) = \eta^2 O(d^2\chi)f(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t)\\
\mathbb{E}[|Q_3 - Q_4|1_{\mathfrak{E}_t}|\mathfrak{F}_t] \le&\eta^3 O(d^4)\sum_{ij}|\tilde{\mathbf{U}}_t\tilde{\mathbf{V}}_t^\top - \mathbf{M}|_{ij}^3 O(\chi)\\
\le&\eta^3 O(d^4\chi)\left\|\tilde{\mathbf{U}}_t\tilde{\mathbf{V}}_t^\top - \mathbf{M}\right\|_\infty f(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t) \le \eta^2 O(d^2\chi)f(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t)\\
\mathbb{E}[|Q_4|1_{\mathfrak{E}_t}|\mathfrak{F}_t] \le&\eta^4 O(d^6)\sum_{ij}(\tilde{\mathbf{U}}_t\tilde{\mathbf{V}}_t^\top - \mathbf{M})_{ij}^4 O(\chi)\\
\le&\eta^4 O(d^6\chi)\left\|\tilde{\mathbf{U}}_t\tilde{\mathbf{V}}_t^\top - \mathbf{M}\right\|_\infty^2 f(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t) \le \eta^2 O(d^2\chi)f(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t)
\end{aligned}$$

This gives in sum that

$$\mathbb{E}[Q_2 1_{\mathfrak{E}_t}|\mathfrak{F}_t] \le \eta^2 O(d^2\chi)f(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t)1_{\mathfrak{E}_t} = \eta^2 O(\mu dk\kappa^2)f(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t)1_{\mathfrak{E}_t}$$

For the second inequality $|Q_1|1_{\mathfrak{E}_t} \le \eta O(\mu dk\kappa^3)f(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t)1_{\mathfrak{E}_t}$ **w.p** 1**:**

$$|Q_1|1_{\mathfrak{E}_t} \le |Q_1 - Q_2|1_{\mathfrak{E}_t} + |Q_2 - Q_3|1_{\mathfrak{E}_t} + |Q_3 - Q_4|1_{\mathfrak{E}_t} + |Q_4|1_{\mathfrak{E}_t}$$

For each term, we can bound as:

$$\begin{aligned}
|Q_1 - Q_2|1_{\mathfrak{E}_t} \le&\eta O(d^2)\left\|\tilde{\mathbf{U}}_t\tilde{\mathbf{V}}_t^\top - \mathbf{M}\right\|_\infty\left\|\tilde{\mathbf{U}}_t\tilde{\mathbf{V}}_t^\top - \mathbf{M}\right\|_F O(\sqrt{\chi}) \le \eta O(d^2\chi\kappa)f(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t)\\
|Q_2 - Q_3|1_{\mathfrak{E}_t} \le&\eta^2 O(d^4)\left\|\tilde{\mathbf{U}}_t\tilde{\mathbf{V}}_t^\top - \mathbf{M}\right\|_\infty^2 O(\chi) \le \eta^2 O(d^4\chi^2\kappa^2)f(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t) = \eta O(d^2\chi\kappa)f(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t)\\
|Q_3 - Q_4|1_{\mathfrak{E}_t} \le&\eta^3 O(d^6)\left\|\tilde{\mathbf{U}}_t\tilde{\mathbf{V}}_t^\top - \mathbf{M}\right\|_\infty^3 O(\chi) \le \eta O(d^2\chi\kappa)f(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t)\\
|Q_4|1_{\mathfrak{E}_t} \le&\eta^4 O(d^8)\left\|\tilde{\mathbf{U}}_t\tilde{\mathbf{V}}_t^\top - \mathbf{M}\right\|_\infty^4 O(\chi) \le \eta O(d^2\chi\kappa)f(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t)
\end{aligned}$$

This gives in sum that, with probability 1:

$$|Q_1|1_{\mathfrak{E}_t} \le \eta O(d^2\chi\kappa)f(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t)1_{\mathfrak{E}_t} = \eta O(\mu dk\kappa^3)f(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t)1_{\mathfrak{E}_t}$$

For the third inequality $\mathbb{E}[Q_1^2 1_{\mathfrak{E}_t}|\mathfrak{F}_t] \leq \eta^2 O(\mu dk\kappa^2) f^2(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t) 1_{\mathfrak{E}_t}$:

$$\mathbb{E}Q_1^2 1_{\mathfrak{E}_t} \leq 4\left[\mathbb{E}(Q_1 - Q_2)^2 1_{\mathfrak{E}_t} + \mathbb{E}(Q_2 - Q_3)^2 1_{\mathfrak{E}_t} + \mathbb{E}(Q_3 - Q_4)^2 1_{\mathfrak{E}_t} + \mathbb{E}Q_4^2 1_{\mathfrak{E}_t}\right]$$

For each term, we can bound as:

$$\begin{aligned}
\mathbb{E}(Q_1 - Q_2)^2 1_{\mathfrak{E}_t} \leq& \eta^2 O(d^2) \sum_{ij} (\tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M})_{ij}^2 \left\|\tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M}\right\|_F^2 O(\chi) \leq \eta^2 O(d^2\chi) f^2(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t) \\
\mathbb{E}(Q_2 - Q_3)^2 1_{\mathfrak{E}_t} \leq& \eta^4 O(d^6) \sum_{ij} (\tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M})_{ij}^4 O(\chi^2) \\
\leq& \eta^4 O(d^6\chi^2) \left\|\tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M}\right\|_\infty^2 \left\|\tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M}\right\|_F^2 \leq \eta^4 O(d^6\chi^3\kappa^2) f^2(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t) \\
\leq& \eta^2 O(d^2\chi) f^2(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t) \\
\mathbb{E}(Q_3 - Q_4)^2 1_{\mathfrak{E}_t} \leq& \eta^6 O(d^{10}) \sum_{ij} |\tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M}|_{ij}^6 O(\chi^2) \\
\leq& \eta^6 O(d^{10}\chi^2) \left\|\tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M}\right\|_\infty^4 \left\|\tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M}\right\|_F^2 \leq \eta^2 O(d^2\chi) f^2(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t) \\
\mathbb{E}Q_4^2 1_{\mathfrak{E}_t} \leq& \eta^8 O(d^{14}) \sum_{ij} (\tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M})_{ij}^8 \delta_{il} O(\chi^2) \\
\leq& \eta^8 O(d^{14}\chi^2) \left\|\tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M}\right\|_\infty^6 \left\|\tilde{\mathbf{U}}_t \tilde{\mathbf{V}}_t^\top - \mathbf{M}\right\|_F^2 \leq \eta^2 O(d^2\chi) f^2(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t)
\end{aligned}$$

This gives in sum that:

$$\mathbb{E}Q_1^2 1_{\mathfrak{E}_t} \leq \eta^2 O(d^2\chi) f^2(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t) 1_{\mathfrak{E}_t} = \eta^2 O(\mu dk\kappa^2) f^2(\tilde{\mathbf{U}}_t, \tilde{\mathbf{V}}_t) 1_{\mathfrak{E}_t}$$

This finishes the proof. □